KAPUST, JEFFRY ALLAN. Parameters Influencing Interobserver Agreement and Observer Accuracy in a Vigilance Analogue to Naturalistic Observation. (1976)
Directed by: Dr. Rosemery O. Nelson. Pp. 108

Interobserver agreement (reliability) is the usual method used to estimate observer accuracy in naturalistic and contrived observations. Despite the warnings by early researchers and the growing interest in methodological problems involved in the observation process, there has been no research explicating the relationship between interobserver agreement and observer accuracy. In addition, there has been little research into the environmental and organismic variables which influence interobserver agreement and observer accuracy.

In an attempt to address these problems, a situation that is analogous to naturalistic observation, namely a vigilance paradigm, was utilized. Experimental assistants performed two arbitrary behaviors (lifting and/or moving the index finger of each hand) at a preprogrammed rate; the behaviors were automatically recorded by electromechanical equipment. In one-hour sessions, the subjects, who were 36 female college undergraduates, recorded the assistant's behaviors by pressing buttons; the subjects' responses were also electromechanically recorded. The experimental design was a two by two by three factorial design with repeated measures across a 60 minute experimental session. Three subjects were nested in each cell. The main factors were: the assistant who was observed (n=2), the distance between the hands of the assistant as she performed the behaviors (n=2: 1 inch or 13 inches), the combination of rates at which the target of observation occurred

(n=3: 1.1 occurrences per minute for both hands; 3 occurrences per minute for both hands; or 1.1 occurrences per minute for the left hand and 3.0 occurrences per minute for the right hand) and 10-minute intervals (n=6). During the observational session, each subject had a button to press when she felt she had made an error in observing or recording.

Multivariate and univariate analyses of variance were performed on observer accuracy data and on the differences between interobserver agreement and observer accuracy. A univariate analysis of variance was performed on error recognition data. The results of these analyses confirmed the two major experimental hypotheses: that interobserver agreement has no direct relationship to observer accuracy, and that observer accuracy is controlled by environmental variables in addition to the targets of observation. The analyses of the accuracy data showed that the accuracy of observation was significantly influenced by the rate of the targets of observation, the number of minutes of ongoing observation, the particular assistant being observed, and the interactions among these and the distance between the targets of observation. Interobserver agreement was found to differ from accuracy of observation by as much as 26%; 7.2% of the total number of differences exceeded 10%. These difference data were also found to vary significantly within the experimental conditions mentioned above.

The implications of these results for the typical experiment utilizing systematic observation are discussed and methods which may reduce or eliminate these problems are presented.

PARAMETERS INFLUENCING INTEROBSERVER AGREEMENT AND OBSERVER ACCURACY IN A VIGILANCE ANALOGUE TO NATURALISTIC OBSERVATION

by

Jeffry Allan Kapust

A Thesis Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
1976

Approved by

Rosemery O. Nelson

Thesis Adviser

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of the Graduate School at the University of North Carolina at Greensboro.

Thesis Adviser Rosemary O. Nelson

Committee Members

3-24-76

Date of Acceptance by Committee

ACKNOWLEDGEMENTS

It is with great appreciation that I thank Dr. Rosemery Nelson for her guidance and support through all phases of this study. I would also like to thank Dr. Russell Harter and Dr. Richard Shull for their numerous helpful suggestions. In addition, I would like to thank Mary Bosch and Carolotta Gabard, the experimental assistants, for their aid throughout this research.

Above all, I would like to express my gratitude to my wife, Debra, for her continual support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Systematic observation procedures involve the systematic quantitative recording of a representative sample of the "behavioral stream" (Barker & Wright, 1951) as it occurs or is seen on film or videotape. One or more observers watch the subject's behavior, assign the behaviors to one or more explicitly defined categories, and manually record the event. In general, systematic observation differs from other types of behavioral assessment on several dimensions. It is performed by a person other than the one displaying the behavior of interest, unlike self-observation and self-rating procedures; it is direct, that is, measures the actual behavior under consideration as it occurs, unlike checklists, interviews, diaries, and questionnaires; and it measures the behaviors under natural contingencies, unlike role playing.

Systematic observation procedures are currently being used by a vast number of therapists and researchers in many fields (Goldfried & Sprafkin, 1974). Systematic observation has been used to record the behavior of many different organisms, including humans, other species of primates, and cetaceans. Investigators in such diverse fields as psychology, sociology, zoology, anthropology, and education use systematic observation to gather data. There are several reasons for this wide use of systematic observation.

First, behaviorism has become accepted in varying degrees by an increasing number of persons working with human and animal behavior. The direct measurement and quantification of behavior are basic tenets of behaviorism (Johnson & Bolstad, 1973; Mash, Terdal & Anderson, 1973). In addition, many behaviorists seek to delineate the environmental events preceeding and following behavior. Systematic observation is ideally suited to these tasks.

Second, systematic observation procedures are inexpensive and versatile (Gellert, 1955). They can be implemented in virtually any situation by trained nonprofessionals (e.g., Eckman, 1973). Any behavior that can be accurately discriminated by the observer can be measured, and several behaviors and individuals can be monitored simultaneously (e.g., Patterson, Ray, Shaw, & Cobb, 1969). Several different systematic observation procedures have been developed and are in common use including time sampling (e.g., O'Leary & Becker, 1967), spot checks (e.g., Kubany & Sloggett, 1973), and continuous monitoring (e.g., Lovaas, Freitag, Gold, & Kassorla, 1965).

Third, systematic observation, especially the use of time sampling procedures, has had a long history of use. However, this long history and the widespread current use of systematic observation procedures does not eliminate certain problems inherent in systematic observation procedures.

Problems of Systematic Observation. Two problems will be examined herein. The first involves the distinction between interobserver agreement and the accuracy of observation. Agreement and accuracy measure different aspects of observation procedures and do not necessarilly permit prediction of one from the other. The second problem involves the degree to which observers' behavior, especially recording behavior, is under the control of stimuli in addition to those they are explicitly observing and recording. (The rationale for these two problems will become apparent below.)

These problems suggest that the model on which systematic observation is based, the traditional reliability model, may not be appropriate. The reliability model was adopted when systematic observation procedures were first developed. A brief history of this development is presented below.

Acceptance of Traditional Reliability Model. Thorne, Schlottman, and Seay (1969) and Arrington (1932) have noted that in the late 1920's several investigators of child behavior, including Arrington, Olson, and Thomas, developed time sampling observation procedures to record the behavior of children. These researchers had become dissatisfied with the then current observation and data collection procedures, which typically were qualitative, unstandardized, and subjective. They viewed the procedures used to measure

personality and intelligence as highly successful and felt that problems in the measurement of childrens behavior could be solved using the testing fields' statistical techniques, models, and terminology (Olson & Cunningham, 1934). Specifically, the child behavior researchers adopted the traditional approach to dealing with variation in measurement: observed score is equal to some true score plus a randomly distributed error, which can be estimated by using various statistical techniques for determining the reliability of measurement. (The adoption of the basic model of reliability was not made explicit in the 1930's, but was a necessary assumption for their use of the word "reliability" and their statistical formula for the calculation of reliability coefficients.)

The child behavior investigators attempted to construct their time sampling procedures to avoid the problems they found in the projective tests, retrospective surveys, and interviews then used (e.g., Arrington, 1932; Olson, 1931). The procedures that they developed (which were the basis of those used today) were objective and non-inferential, and involved quantitative, precise, and non-interpretive scoring systems. In accordance with the tenets of testing theory, the investigators attempted to determine the extent of measurement error in their time sampling procedures (i.e., reliability). Their measure, which they called interobserver reliability (Gellert, 1955), consisted of the comparison of the simultaneous observations made by two or more independent observers (Arrington, 1932; Olson, 1931).

Current investigators still attempt to determine the measurement error in their observation procedures. Determination of interobserver reliability has become a requirement for any experiment utilizing any form of systematic observation (e.g., Preparation of Manuscripts for the Journal of Applied Behavior Analysis, 1969). High levels of agreement between observers are assumed to reflect a highly reliable measurement procedure. Repp, Deitz, Boles, Deitz, and Repp (in Press) state that while some authors refer to this measure as observer reliability, others use the term observer agreement or interobserver agreement. Johnson and Bolstad (1973) present a discussion of these terms. From this point on, interobserver agreement will be used to refer to the measurement obtained through simultaneous recording by two or more independent observers. The reliability concept will be shown to be inappropriate for the above measurement situation.

Measures of Reliability. The traditional testing definition of reliability is the degree to which a device produces similar results on subsequent or simultaneous measurements of an event or object under the same conditions (Gellert, 1955). In other words, reliability is the extent of agreement between measures when the measurement conditions are maximally similar (Campbell & Fiske, 1969). Thus, reliability is the ability of the tool (procedure) to produce replicable results.

Several types of reliability have been developed and used. They include test-retest, internal consistency (split-half), and alternate or parallel forms reliability. Each one is concerned with a different aspect of the measurement process. Kelley (Chapter 4, 1967) gives the following definitions of the three types of reliability. Test-retest reliability involves comparing the results of two temporally separated administrations of the same measurement device. Split-half reliability measures the internal consistency of the measurement tool. It can only be used when the tool consists of a number of items or units, each contributing equally to the total measurement. The test is divided into equal portions (usually using the odd numbered items in one half and the even numbered items in the other) and the two half measurements produced through administration of the overall test are compared. The reliability score that is computed from the two half-tests is then adjusted for the attenuation caused by splitting the original test. The third type of reliability, alternate or parallel forms reliability, involves the comparison of scores produced by two or more equivalent measurement devices.

All three types of reliability have been applied to systematic observation procedures. Johnson and Bolstad (1973) and Jones, Reid, and Patterson (1975) provide detailed discussions of these applications. Test-retest reliability is used to

estimate the stability of subjects' behavior across time when the situations in which the behaviors occur are constant. Split-half reliability is used to determine the consistency of observation across several days or weeks. The odd and even days of observation are compared to each other. The observation setting is not altered throughout the observations. Alternate form reliability is the method most often used with systematic observation procedures. It is directly related to interobserver agreement (observer reliability) both in theory and computational procedure. Each observer is considered to be one form of the measurement tool and to be an equivalent form of the other (Jones et al., 1975; Olson & Cunningham, 1934).

Accuracy of Observation. The ultimate goal of systematic observation is to produce a record of behavior that is as consistent as possible with the actual behavior. The accuracy of a measurement reflects the extent to which the measurement represents some dimensions of the entity being measured. In systematic observation, these dimensions range from relatively simple (physical extent of movement and frequency of occurrence) to very complex (interactions between topographical, spatial, and temporal characteristics).

Determination of the accuracy of observation requires knowledge of the actual event being observed. This information is quantitatively compared to the data obtained through observation. Accuracy of the observation is inversely related to the magnitude of the difference between these quantities.

However, in the usual naturalistic observation situation, this knowledge of the actual event is unavailable (observation is the method by which the event is measured). In order to circumvent this problem in studies of the observation process, current investigators have experimentally defined the actual event being measured through the use of observer protocols or standard situations (e.g., Johnson & Bolstad, 1973).

The concept of reliability of measurement is different from the concept of accuracy. Reliability indicates the consistency of measurement; the degree to which a measurement tool produces the same information in the same situation. Reliability makes no statement about the degree to which the information is in accordance with the actual event that is being measured. For example, a poorly adjusted speedometer may be extremely innaccurate: read 65 MPH when the actual speed is 70 MPH. However, it may be extremely reliable and always read 65 MPH when the actual speed is 70 MPH.

Interobserver agreement, used to estimate the reliability of measurement, also relates to accuracy in the above manner: it permits no inferences to be made with respect to the actual accuracy of observation. However, the widespread use of interobserver agreement and the neglect of the accuracy of observation in virtually all research using systematic observation suggests that researchers may be erroneously using interobserver agreement as a measure of the accuracy of observation.

Problems with Reliability Model. Many researchers have qualified their acceptance of the reliability model and the use of interobserver agreement to estimate reliability (e.g., Gellert, 1955; Olson & Cunningham, 1934). These qualifications usually involved the listing of variables that were thought to influence reliability and interobserver agreement (e.g., number of subjects that are simultaneously observed). However, until the late 1960's there was no direct empirical evidence for these variables. In recent years, many researchers have begun to investigate some of the hypothesized factors (e.g., Bobbitt, Gordan & Jenson, 1966; Johnson & Bolstad, 1973; Reid, 1970), yet even these researchers have limited themselves to variables within the reliability framework.

However, with respect to naturalistic observation, the reliability model itself may be subject to question. There are several problems that arise from this use of the various types of reliability and the determination of interobserver agreement. One problem stems from the inability of the researcher to separate the obtained data into variation due to changes in subjects behavior and variation due to an unreliable measurement tool. Test-retest and split-half reliability both suffer from this problem. Even though the experimenter may attempt to control the situation in which the behavior being observed occurs, there is no guarantee that the behavior does not change. Thus, consist ncy of the measurement tool and consist ncy of the measured entity are confounded in these two types of reliabilities.

Because of the above problem with test-retest and split-half reliabilities, alternate forms reliability has been the most widely used procedure. However, it is the basic purpose of this paper to demonstrate that observers are not equivalent forms of each other.

Experimental Hypotheses. Two experimental hypotheses are necessary to fulfill the above purpose. The first states that interobserver agreement does not predict the accuracy of the observers. Thus, one or more observers may be performing at substandard levels, while agreement between observers is quite high. The second hypothesis suggests the cause of the first. It states that an observer's accuracy is influenced both by stimuli related to the behavior being observed and by other unrelated stimuli in an individual manner. That is, the accuracy of a set of observations, and thus the degree to which interobserver agreement deviates from this accuracy, changes as a function of various aspects of the subjects' behavior, other environmental stimuli, and methodological variables. The typical conceptualization of variables which control the observer's behavior, i.e., the reliability model, includes only those behaviors of the subject to which the experimenter has directed the observers' attention. This conceptualization does not account for possible control of the observers' behavior by other variables, nor for possible interactions between variables and the specific observer.

If these latter variables do exert control over the observers, the traditional model of observers as unchanging, equivalent forms of each other is untenable.

Evidence for the Hypotheses. Two large and growing bodies of evidence against the parallel forms conception of the observer and the use of interobserver agreement as an estimate of the reliability of observation are provided by research into observational methodology and vigilance research.

The first body of evidence for the most part has appeared quite recently. It includes research which demonstrates the influence that methodological variables and stimuli that are incidental to the targets of observation have on interobserver agreement.

Data from an area labeled observer reactivity (Lipinski & Nelson, 1974) demonstrate that the agreement between observers is a function of the manner (or even existence) of interobserver agreement assessment. Covert assessment procedures have repeatedly shown lower agreement than overt procedures (Reid, 1970; Taplin & Reid, 1973). Overt assessment on one day does not necessarilly reflect levels of agreement on other days. Also, the presence of specific observer partners is sufficient to change the observers' behavior in ways that result in higher agreement (e.g., Romanczyk, et al., 1973). In addition to observer reactivity, factors labeled observer bias and observee reactivity, and various methodological

problems involved in the agreement assessment process contribute to observational inaccuracy. For discussion of these factors, see Johnson and Bolstad (1973) and Jones, et al. (1975).

Additional findings to be described below demonstrate that the behavior of observers is affected in different ways by the following variables: complexity (difficulty), rate, and predictability of the behavior being observed, number of categories, and background information about the subjects. Gellert (1955) stated that there is a need for assessing interobserver agreement in both difficult and easy recording situations. The ease of recording is directly related to such variables as the frequency of each behavior, changing topographies, and the predictability of the behaviors. Johnson and Bolstad (1973), Jones, et al. (1975) and Reid and Skindrud (1973) have also voiced this need, saying that spuriously high or low coefficients of agreement may be calculated if this is not done. King, Ehrmann, and Johnston (1952) demonstrated that interobserver agreement was reduced when identical information about the subjects was supplied to the observers. Reid, Skindrud, Taplin and Jones (1973) have suggested that an inverse relationship exists between interobserver agreement and the complexity of the recorded behavior. Many investigators (e.g., Bobbitt, Gordan, & Jenson, 1966; Patterson, cited in Bolstad & Johnson, 1972; Thorne, Schlottman, & Seay, 1969) have stated that as the frequency of a response decreases,

reliability as estimated by interobserver agreement also decreases. Mash & McElwee (1974) have reported that the predictability of behavior, the number and complexity of catagories being simultaneously observed, and the observer's history of observing predictable behavior all influence observer agreement. They found that eight catagories produce lower agreement and lower absolute accuracy than four; predictable behaviors yield higher agreement and accuracy than unpredictable or previously predictable behaviors.

The evidence cited above shows that the procedures used to calculate and measure interobserver agreement influence the score which is produced. Also, it shows that identical stimuli can influence different observers' behaviors (including the acts of observing and recording behavior) in different manners.

The second body of data has existed since the 1940's in the area of vigilance. Vigilance studies typically investigate the variables which influence the ability of an observer to recognize and respond to a signal that occurs in the environment. The typical vigilance paradigm involves an observation task in which the signal can be discriminated from the non-signal (a moderate signal-to-noise ratio), the probability of a signal is very low (about .1/minute), and the duration of the observation period is long. The task itself can be visual (monitoring a radar scope) or auditory (monitoring sonar). There is either a periodic or continuous

presentation of a display of stimuli which may or may not contain the signal. The frequency of these presentations is called the event rate. Thus, specific events are presented to the observer at a predetermined rate. At predetermined times, a signal is substituted for or combined with the event. The observer must detect and respond to these occasional signals.

The subject (observer) in a vigilance experiment is assumed to emit three responses. The first, the observing response, consists of a sequence of behaviors which permit important environmental stimuli (including the signal, if it is present) to be received by the observer. The second response, the decision, is a usually covert comparison of the perceived information to a usually internalized representation of the signal. If the observer classifies the observed information as a signal, a specific response must be emitted. If the observed information is classified as a non-signal, the observer may or may not be required to emit a specific response, depending upon the experimental requirements. This, the recording response, is the third portion of the vigilance task. The relevant vigilance data can be divided into three general sections: procedural (methodological) variables, parameters of non-signal stimuli and parameters of the signal.

Many procedural variables which have been manipulated in vigilance experiments have been found to increase accuracy and maintain it at high levels (McGrath, et al., 1959). However, two of these variables, knowledge of results and the

use of consequent stimuli (reinforcers and punishers), are not applicable to the typical systematic observation situation due to the experimenter's lack of knowledge about the actual occurrence of the signal in these settings.

When an observer is presented with information about his performance either as feedback, per se, or in the form of rewards or punishments, his accuracy increases and is maintained. The information or consequences may be presented after every recording response and missed signal or at regular intervals throughout the experiment.

One of these procedural variables from the vigilance field that does appear to be valuable for the observation field is the frequent interpolation of rest (non-recording) periods. This appears to maintain accuracy at higher and more stable levels than constant recording. In contrast, continuous recording appears, under certain conditions, to cause a decrease in accuracy over time which reaches an asymptote after approximately one half hour of observation. This decrease has been labeled the performance decrement. In addition, latency (the interval between the signal presentation and the observer's response to the signal), which usually increases across time, is maintained at low levels with rest periods. These findings are relevant to the decision to use continuous versus discontinuous observation procedures and raise questions about the relative accuracies

of on-off types of recording, such as the O'Leary code, time-sampling, and spot checks in comparison to those of the continuous types, such as the Patterson Family Interaction Code or frequency counts. Another possibly useful procedural variable, the degree of complexity of the observation task (number of displays being observed and the number of responses to each display) affects accuracy in a complex manner. Increased task complexity apparently eliminates the performance decrement, but lowers overall signal detection levels appreciably. Other parameters of the observation situation may interact with the procedural variables. There is some evidence for this interaction (see below) but definitive research is lacking.

A second general area in vigilance research, with greater implications for the parallel forms concept than the procedural area, is comprised of the parameters of the non-signal stimuli impinging upon the observer. These variables involve environmental stimuli which, though not a portion of the stimulus display, are perceived by the observer and influence his behavior. Few aspects of this area which are not specific to the display have been studied, but the ones which have, demonstrate this area's importance. Time of day appears to influence performance in an unsystematic manner which is specific to the individual (Jenkins, 1958). Overall noise level in the vigilance area (which in some systematic

observation situations is directly related to the rate and complexity of the behaviors being observed) reduces detection performance in complex tasks but not in simple ones (Broadbent, 1954; Loeb & Jeantheau, 1958). In addition, there appears to be an optimal temperature range for observing (Mackworth, 1950). The length of the recording (vigilance) session is directly related to overall accuracy. Performance usually decreases as time on task increases to about 1/2 hour after which the detection level is constant, but low (the performance decrement). This finding occurs only in certain combinations of signal rate and complexity, and task complexity.

The last area of vigilance evidence involves parameters of the signal and of the observing response. Typically, the signal in vigilance experiments has been of high intensity and very low frequency. The event rate, which determines the rate at which observing responses must be made, is usually quite high (1 or .5/sec.). These conditions are analogous to those in typical radar or sonar watches. However, other rates have been investigated.

The first parameter of the signal is rate. Most researchers report that performance increases as the absolute signal rate increases. In addition, at higher signal rates the performance decrement is eliminated. However, recently, Jerison and others, using an observing response and reinforcement model, have demonstrated that the event rate controls the level of detection performance and the existance of the performance decrement independent of the absolute signal rate.

The ratio of the signal and the event rates is apparently the controlling factor. In addition, they have shown that the latency of the detection response is inversely related to event rate. Good detection performance is accompanied by long latencies; poor performance, by short response latencies. These data reveal that the rate at which observing responses are elicited determines the observer's performance. Before a signal can be detected and appropriately responded to, the observing response must be emitted. As the event rate (the frequency of possible signals) increases, the observing response rate must increase to maintain the observer's accuracy. If the signal rate is not increased proportionally, the probability that a particular observing response will be reinforced by the detection of a signal decreases. At the very low reinforcement rates typically found in the vigilance experiment, the observing response begins to extinguish. The initial high level of responding may be maintained by internal factors, such as military orders, desire to please the experimenter, and false expectations (see below). As fewer observing responses are emitted, the probability of a detection and, thus of a reinforcement, decreases. This reduction of reinforcement increases the extinction process and causes the typical performance decrement. Latency is related to the event rate in the following manner: as the event rate increases, there is a shorter period in which the observing response-decision-recording response sequence can

occur. As the observing response extinguishes, more observations are made in which the signal/non-signal decision is not clear cut. This condition may be caused by changes in the topography of the observing response. When the observer is given enough time to make the decision (which always occurs when a low event rate is presented and which can occur during a fast rate condition if the observer ignores subsequent presentation of events while he is deciding), he makes few mistakes. As he is forced to make the decisions at a faster rate (which produces a shorter latency), the quality of the decision, and thus the accuracy, decreases. The theory does not account for the eventual asymptotic level of the performance decrement and for the absence of the decrement when certain variables are manipulated (see below). Additional evidence for the observing response theory is found when factors which influence and control responding during extinction are investigated. Two of these, interpolated rest periods and presentation of artificial signals, which decrease the course of extinction, maintain vigilance performance at high levels. The overall rate rather than particular schedules or inter-signal interval is most commonly investigated because the latter variables have produced results which are uncertain due to confounding of the schedule or inter-signal interval with the event rate. However, McGrath et al. (1959) report

that the more variable the schedule the lower the detection performance.

Variability is directly related to the second parameter of the signal: temporal uncertainty and a priori knowledge of the signal probability. Adams & Boulter (1964) found that a predictable signal yields higher detection performance than an unpredictable signal. However, Braddeley & Calquhoun (1969) have shown that the typical procedure in vigilance experiments which gives the subject a practice session with a signal rate that is much higher than that in the main experiment creates a spuriously high expectation of signal rate. This false expectation may cause the results of the typical vigilance study. When appropriate expectations are established (as in the above experiment), the performance decrement occurs only when extremely low signal rates are presented. Experimental manipulation of pre-observation expectancies yields data which are consonant with the extinction theory of vigilance. That is, even when the subject is informed of (given practice with) the low signal rate, the observing response extinguishes when the signal to event rate is very low. The expectation matching procedure may remove a cognitive factor which expediates extinction.

The third signal parameter which has been investigated is spatial location and predictability. Unfortunately, changes in location have not been investigated independently

of spatial uncertainty and, in some cases, task complexity and signal and event rates. While these factors limit the interpretation of these experiments, the general findings are that spatial uncertainty of one stimulus decreases detection performance (Adams & Boulter, 1964) and increases reaction time. When several stimuli occur at different locations (e.g., Loeb & Alluisi, 1970), the task complexity and signal-event rate ratio considerations render the results uninterpretable.

Other aspects of the signal which appear to influence performance and which are relevant to systematic observation are intensity and duration. Loeb & Alluisi (1970) also report that as signal duration decreases, the subject's performance decreases. As signal duration increases, the upper limit of the relationship appears to be two to four seconds. As the intensity of the signal increases from low to moderate levels, the detection performance increases and reaction times decrease.

The above data are usually reported as group means. If the group variances for these, and for presently uninvestigated, variables are low, that is, if most observers react to the same stimuli consistently in the same manner, the variables would be of little relevance to systematic observation. The extent of any observational error could be determined through extensive research. Correction factors for

particular techniques and situations could be developed and applied to the data. However, McGrath et al. (1959) report that individual differences in vigilance performance are often quite large, especially with respect to certain variables. Taub & Osborne (1968) have demonstrated that error rate increases with observation time, but that there is large subject variability. In addition, several investigators (Holland, 1958; Mackworth, 1950; Solandt & Partridge, 1946) have suggested that 'expert observers' exist. These persons do not show the performance decrement and are more accurate in more situations than the 'typical' subject. Estimates of the number of 'expert' observers in the general population range from 20% - 50%. This aspect of vigilance has not been investigated in detail, but it appears that the typical vigilance performance is attributable to only some group members.

The data from the observational methodology and the vigilance areas show the degree to which various procedural variables influence the magnitude of both interobserver agreement and observer accuracy. These data also demonstrate that stimuli found in the typical observation situation which are not direct targets of observation can profoundly alter interobserver agreement and accuracy. In addition, there is evidence that observers are influenced in different ways by stimuli (both target and non-target) that are identical.

Thus, the observer in the naturalistic observation situation is not a reliable measurement tool. Evidence has been presented which demonstrates that the instrument (the observer) can change within and between measurement sessions. The observer is also extremely sensitive but not only to target stimuli. Many variables control the observer's behavior but few of their effects have been experimentally determined. Thus, observers are not parallel forms of each other. Each observer is influenced by internal and external stimuli in a manner that is usually similar, but not identical to all other observers.

These results are in agreement with those of Endler and Hunt (1966) who, in an investigation of the parameters of the verbal report of anxiety, found that the response mode, the stimulus situation, individual differences, and the interactions between them all contributed variance to their results. In the systematic observation experiment, the method of observation is the response mode; all environmental stimuli--including those important to the investigator as well as others--are the stimulus situation. Both of these and the individuality of the observer interact to influence the accuracy of observation.

Statement of the Problem. As the situation now stands, evidence exists which questions the current use of interobserver agreement to estimate observer accuracy and which casts doubt upon the continued use of the reliability model in the context of interobserver agreement. In addition, a

large body of data supports the hypothesis that observers' behavior is not only under the control of the targets of the observation but is also influenced by other variables. However, there is currently no evidence which can definitively state the degree to which of the above hypotheses are supported; that is, which specifies the exact relationship between interobserver agreement and observer accuracy and indicates the manner in which various variables influence accuracy.

To produce this type of evidence, several experimental conditions must be met. First, the observation setting must be simple to facilitate interpretation and avoid experimental confounds. Second, the observation conditions must be under complete experimental control, yet they must also be analogous to many 'everyday' observation situations to maintain the generalizability of the results. Third, both the stimuli which are intended to elicit the observers' responses (the behavior of the subjects) and the behavior of the observers (the recording responses) must be simultaneously recorded. In this way, comparisons can be made between the actual event, and the observers' recordings of the event, as well as comparisons between the observers' recordings. In this way, the various accuracies, interobserver agreements, and error types can be determined and compared.

The vigilance paradigm permits the above conditions to be fulfilled and is a subset of systematic observation. However, certain aspects of the vigilance experiment must be altered to permit generalization to naturalistic observation situations. Although the vigilance experiment does provide information about actual signal parameters (rate, intensity, predictability, etc.), the parameters of the signal, the recording response, and other variables that are used in typical vigilance research have differed greatly from those found in the typical naturalistic observation situation. In the natural environment, many variables, such as signal location, topography, duration, are not controlled and change during the course of a single observation. Also, systematic observation often involves the simultaneous observation of several behaviors.

The present experiment utilized the vigilance model, incorporating parameters that were typical of systematic observation. This vigilance anologe of systematic observation permitted the accurate monitoring of all data yet permitted the use of desired parameters.

Twelve groups of three subjects were required to observe and record the behavior of two assistants to the experimenter. Six groups observed each assistant. Two arbitrarily chosen behaviors both involving movement of the assistants' hands were observed. The behaviors were considered

typical of those involved in systematic observation in rate, intensity, duration, predictability, etc. The subjects observed the assistant for one 60-minute session. The subjects' observations and assistants' behaviors were recorded simultaneously. This permitted quantitative analysis of each subject's accuracy within each observation session. Interobserver agreement was calculated for each pair of subjects (three agreements for each session) and compared to the actual occurrence of the behaviors (the subject's accuracies). Each subject was provided with a means of recording any errors she felt she made. These records were compared to the subjects' actual accuracy. In addition, comparisons of the above data were made between subjects exposed to different rate and spatial separation combinations. There were three rate conditions: both behaviors being observed occurring at High rate (three signals per minute); one behavior at High rate and the other behavior at Low rate (1.1 signal per minute); and both behaviors at Low rate. There were two spatial separation conditions: one in which the observed behaviors were separated by one inch (the Together condition) and one in which there was a 13 inch separation (the Separate condition). Rate, Spatial Separation and the particular assistant being observed were factorally combined. All other parameters (topography, intensity, etc.) were held constant across time and between groups.

In addition to the two major hypotheses discussed above, the following specific outcomes for accuracy data were hypothesized from pilot data and vigilance research:

1) There would be no main effects due to differences in the presentation of stimuli between the two assistants.

2) There would be a main effect of Spatial Separation, with the Together condition showing higher overall accuracy than the Separate condition.

3) The effects of the Rate parameter were not aboe to be predicted from pilot data. However, an interaction between Rate and Spatial Separation was expected. This interaction was predicted to occur in the High-Low Rate condition, taking the form of increased accuracy on the fast behavior or decreased accuracy on the slow behavior when the behaviors were separated but no change when they were not separated. In addition, the Low-Low Rate condition was thought to yield lower accuracy than the High-High Rate condition. This was predicted by Jerison's theory of observing response extinction (discussed above).

4) Change in accuracy across time was expected from most subjects.

5) Large individual differences in observers' accuracies were expected.

CHAPTER II

METHOD

Subjects and Assistants

Thirty-six female undergraduates who reported that their visual acuity was corrected to at least 20-20 were randomly selected from the Human Subjects Pool at the University of North Carolina at Greensboro. All subjects who attended the experiment received credit in Introductory Psychology.

Two female undergraduates served as assistants to the experimenter. Their performance in this experiment was a portion of an independent study project. Both assistants were slim, attractive, and approximately 66 to 68 inches tall. Both appeared highly motivated and performed extremely boring tasks carefully and cheerfully throughout the experiment. Both assistants were majors in psychology and graduated following the end of the experiment. The assistants' responses to programmed cues served as the stimuli which were observed by the subjects. The assistants also assisted in certain portions of the data analysis.

Observation Task and Definition of Target Behaviors

The subjects were required to observe an assistant who was moving her index fingers above and on metal touchplates in a programmed manner. The subjects recorded their observations by pressing pushbuttons.

There were two touchplates for each index finger; one located in front of the other. The position of the right index finger that was the target of observation was touching the right, front touchplate (i.e., that touchplate that was closest to the subjects and furthest from the assistant). The target position for the left index finger was touching the left, back touchplate (i.e., that touchplate which was closest to the assistant and furthest from the subjects). The assistant's finger movements were limited to the following: lifting the index finger from the touchplate on which it was resting and holding it 3/4 inches above that or the other touchplate; moving the finger from 3/4 inches above one touchplate to 3/4 inches above the other touchplate; moving the finger from 3/4 inches above one touchplate to touching that or the other touchplate, and moving the finger from touching one touchplate to touching the other touchplate. Each finger moved between one front-back pair of touchplates only.

The assistant's other fingers were kept in a fist position (touching the palm). Her forearms and the heels of her palms rested against the table top during observation periods and were moved only as the index fingers moved forward and backward.

The subjects were required to press one pushbutton whenever the left index finger assumed its target position (i.e., touched the left, back touchplate) and continue pressing

until the finger moved from the target position. A second pushbutton was used, in the same manner, to record observations of the right index finger in its target position.

Design

The experiment had a two by two by three factorial design (between) with repeated measures across a 60 minute experimental session. Three subjects were nested within each of the 12 cells. The main factors were: the particular assistant whose behaviors served as stimuli (two levels); the spatial separation of the observed behaviors (two levels); the scheduled rate of cue presentation (three levels); and, the 10-minute interval of time from which the measurements were taken (six levels).

To control for possible differences in the presentation of stimuli, two assistants were observed, each by one half of the subjects. Spatial separation, the second main factor, was the distance between the assistant's index finger during observation. One half of the subjects observed the assistants' fingers in the separate condition (left and right pairs of touchplates separated by 13 inches), while one half observed the assistants' fingers in the together condition (left and right touchplates separated by 1 inch). These particular distances were selected so that the assistants were able to comfortably perform their task and the subjects were able to clearly distinguish between the touchplates in the together

condition and unable to accurately percieve both fingers in any single fixation in the separate condition.

The rate factor involved the rate at which the target positions were assumed by the assistant's fingers. Each finger changed position at a rate of 8.1 movements (position changes) per minute. This was the event rate. The target position was assumed (changed to from another position) either 3.0 times per minute (High rate) or 1.1 times per minute (Low rate). These rates were the signal rates. Thus, 8.1 times per minute an event (a position change) occurred. In a High signal rate condition, 3.0 of the events were changes to the target position for that hand. In a Low signal rate condition, 1.1 of the events were changes to the target position. In both conditions, the remaining events (position changes) consisted of changes to and among the three non-target positions for each finger.

Because the present experiment was designed as an analogue to systematic observation, signal rates were selected that were greater that those typical of vigilance research (usually about .1/min.). Conversely, an event rate was selected which was less than those typical of vigilance research (usually about 60/min.). The signal and event rates selected were thought to be typical of those often found in research that utilizes systematic observation.

The three levels of the Rate factor were determined by the signal rates for each hand. The levels were: High-High, in which both index fingers assumed the target position 3.0

times/minute; Low-Low, in which both index fingers assumed the target position 1.1 times/minute; and Low-High, in which the assistant's left index finger assumed its target position at the Low rate (1.1 times/min.) while the right index finger did so at the High rate (3.0 times/min.). The fourth possible rate combination (High-Low) was not included because it was assumed that there was no topographical differences between the target positions for the index fingers.

The order of the positions to which the assistants moved their index fingers was randomly determined with the above rate constraints. Four lights served to signal the assistant to change positions, where one pair of lights corresponded to each index finger. The lights were controlled by electronic programming equipment. One light of each pair signalled a new event (i.e., it indicated that a new position was to be assumed). The other light of each pair indicated that the next event was to be the target event for that index finger. Due to methodological difficulties, the assistants determined which of the three non-target positions were to be assumed when a move to a non-target position was appropriate. The assistants were instructed to assume non-target positions in as random an order as was possible. The programmed sequence of events and target positions was 10 minutes long and was repeated six times over the 60 minute observation period. Thus, in each session, the same set of stimuli were presented to the subjects six times. These six time intervals served as the fourth factor in the design.

Apparatus

Each subject was seated before a modified Lafayette 632AS Visual Choice Reaction Time Apparatus. The basic device consisted of a horizontal row of four lights of different colors parallel to a row of four pushbuttons. Each pushbutton was adjacent to a light. The pushbutton and light on the far right of the device were covered with black tape and were not used in the experiment. Labels were placed between the pushbutton and the light of the three remaining pairs and served to identify the function of each pushbutton. The subjects used the two left hand pushbuttons to record their observations of the assistant's fingers. The subjects were instructed to press the far left pushbutton when the assistant's right index finger assumed its target position and to continue pressing until the finger was moved to another position. Similarly, pressing the second pushbutton from the left corresponded to touching the target touchplate for the left index finger. The subjects were instructed to briefly press the remaining pushbutton to indicate that she felt she had made an error in observing or recording.

Depression of each pushbutton illuminated the light adjacent to the pushbutton and deflected a pen on anEsterline-Angus Operation Recorder, Model 620A. The pen returned to its resting position when the pushbutton was released.

The assistant sat facing the three subjects who were observing her. A panel containing a single row of four lights was placed in front of the assistant. These lights were not visible to the subjects, nor did the panel obscure any subject's view of the assistant's hands. The lights served as cues to the assistant to appropriately position her fingers. Punched tapes on four 16mm tape readers controlled the cue lights; one tape reader controlled each light. Two tapes were needed to provide cues for each finger; one to control the light that cued events (assumption of a position) and one to control the light that cued signals (assumption of the target position). Both event tapes were identical to ensure the presentation of an equal number of events from each index finger. The signal tapes differed in the rate at which the signals were presented. Because the experimental design contained conditions in which both index fingers presented signals at the same rate, two identical signal tapes were made for each rate condition. All tapes were loops exactly 60cm long (10 minutes).

During the low-high rate condition, one signal tape of each rate was placed on the two tape readers that controlled the signal cue lights. During the high-high and low-low rate conditions, both signal tapes were of the same rate. However, to ensure that there was no predictable pattern between the two fingers, while retaining equal rates of presentation, the signal and event tapes for one hand were reversed on the tape

readers so they were read beginning at the opposite end of the signal and event tapes for the other hand (i.e., they were read backwards).

The subjects sat at a table in a row approximately two feet from each other. They faced a table and chair at the other side of the room (approximately 20 feet away) where the assistant sat. The subjects' recording sets were placed on the table in front of them. They were prevented from seeing each other's recording set by partitions placed between them. All three subjects were able to see both of the assistant's hands at all times.

In order to compare the programmed changes in the stimulus lights to the assistant's actual behaviors (and to subsequently compare the assistant's behavior to the subjects' recordings), metal sewing thimbles were placed on the end of the first joint of the assistant's left and right index fingers (yellow rubber laundry gloves were worn to prevent shock). Six 1.5 inch square metal touchplates were nailed to a wooden board which was placed on the table in front of the assistant. The touchplates were arranged in two identical, parallel rows of three plates. The rows were separated by 1 inch and were parallel to the assistant's body plane. The left hand touchplate in each row was located 13 inches from the middle touchplate which was separated from the right touchplate by 1 inch. Thus, the outside edges of the two rows of touchplates formed a

rectangle 4 inches by 16.5 inches. Only four of the six touchplates were used during any experimental session. The pairs on the far left and right were used during separate conditions; the two right hand pairs were used during together conditions. The unused touchplates were not moved or covered in any way. The fingertip electrodes were connected to the negative output of a 12 volt power supply. The positive output was connected to ground in the programming equipment. Each touchplate was connected to solid state programming which controlled a pen on the operation recorder. Thus, touching a fingertip to the touchplate on the table caused the appropriate pen to deflect until the finger was removed.

Procedure

Pilot Experiment. A pilot experiment was run during October, 1974. In this experiment, two subjects were placed in each group. The signal rate in the low rate condition was two per minute. One touchplate was used for each hand and there was only one non-target position (holding the appropriate finger 3/4 inches above the touchplate). In addition, the index and middle fingers of the right hand were used during the together condition, rather than both index fingers as in the present experiment.

Twelve subjects were used in the experiment. All were given the high-low rate condition; half were exposed to the separate condition and half to the together condition. Three pairs of subjects were given two sessions and three pairs were given one.

Preliminary results showed that 1) there were large individual differences in accuracy of recording, 2) inter-observer agreement did not consistently reflect each or both subjects' accuracy, 3) some subjects could attain high accuracy on the task (as high as .96 - .97) with most response latencies less than about one half second, and 4) there were no systematic changes across time within a session.

The pilot data suggested that the more accurate observers were performing at a ceiling. Therefore, following the pilot experiment, the following changes (previously described in Method) were made in the design in order to reduce the ceiling effect. The low rate condition was reduced from two occurrences of the target position per minute to 1.1 per minute; spatial certainty (the degree to which the location of a behavior can be predicted) was reduced from one target position and one non-target position to one target and three non-target positions. These changes were instituted to increase the number and magnitude of errors in the together condition. Also, the lower target-to-non-target ratio was assumed to be more like that found in naturalistic systematic observation. In addition, in the main experiment, the assistants used only their left and right index fingers in both the separate and together conditions rather than using the left and right index fingers in the separate condition and the right index and middle fingers in the together condition as was done in the pilot study.

Assistant Training. The first phase of the experiment involved training the assistants to move to the appropriate positions on the touchplates according to the cue lights. During this phase, both the cues for the assistant's movements and the assistant's responses to the cues were recorded on the operation recorder. This recording permitted the experimenter to monitor the assistant's error rate and her response latencies. No subjects were present during these sessions. After each session, the experimenter provided feedback about the assistant's performance and all problems were discussed and resolved. Both assistants were required to meet the criteria of no response latencies greater than two seconds from cue presentation and no more than four errors per session of 60 minutes (an error was defined as an omission of a behavior state change or a state change added to or substituted for another). Overt assessment of the assistants' performances in the following of cue light schedules during the data collection phase was made and explicit feedback was given after every session. It was assumed that this procedure would maintain the assistants' performances above criterion level. Neither assistant fell below criterion during data collection, although each made at least one error during each session.

Data Collection. While the assistants were being trained, 36 subjects meeting the criteria specified for subject selection were selected and randomly assigned to 12 groups of three. These groups were then randomly assigned to the 12 cells of the experiment.

The data collection phase took place during the first three weeks of February and was fifteen days in length. For ease of scheduling subjects, each assistant was observed only three times per week. It was planned that both assistants would be observed each of the three days but several no-show subjects, the sickness of one assistant and persistent scheduling problems forced sessions to be held whenever feasible. The order of the groups for Assistant 1 was as follows: left at Low rate, right at High rate, separate; both left and right at High rate, separate; both left and right at Low rate, separate; left at Low rate, right at High rate, together; both at High rate, together; both at Low rate, together. The order of the groups for Assistant 2 was as follows: left at Low rate, right at High rate, separate; left at Low rate, right at High rate, together; both at Low rate, separate; both at Low rate, together; both at High rate, separate; both at High rate, together. On four days, the groups for each assistant were run singly and on four other days, one assistant was run after the other. On the latter 4 days, each assistant was run first two times. Neither assistant was present when the other was being observed nor were any subjects not scheduled to observe present during any observation session. While the experiment was in progress, each subject had a 3 x 5 index card in front of her with the appropriate target positions written on it to aid her if she became confused or forgot the appropriate responses.

Each session involved a 5 minute orientation period, 5 minutes of practice observation, a short question period, 5 more minutes of practice observation, 60 minutes of data collection, then a short debriefing period. The total time required of each subject was approximately 80 minutes. Each subject observed for only one session.

The following is the procedure that was followed in all observation sessions. The experimenter and the scheduled assistant set up the experimental apparatus and ensured that it was functioning properly. The assistant then donned the rubber gloves, fastened electrodes (thimbles) to her fingers and ensured that all circuits were fully operational. When the three scheduled subjects arrived, orientation began. If less than three subjects were present 15 minutes after the scheduled starting time, the session was cancelled and the subjects present were rescheduled. No no-show subjects were rescheduled. (Three sessions were rescheduled due to no-show subjects.) Orientation consisted of visual acuity screening using an Armed Forces Visual Acuity Test, Form 3 (which is very similar to the Standard Snellen Chart) followed by a general statement consisting of the purpose of the experiment and the experimental task and targets. Only two subjects scored less than 20/20 on the vision test; these subjects both scored 20/25. After the orientation, a five minute practice session was held. In this session, the subjects familiarized themselves with the recording set and attempted to record the assistant's behaviors. All questions or problems

that arose were fully answered. A second 5 minute practice session was then held. The last portion of the observation session was the data collection period which lasted 60 minutes. The experimenter read the following statement during the session:

> You are about to participate in an experiment which will require about 80 minutes of extreme concentration and attention. Anyone not willing or able to give this attention, should withdraw now. Does everyone wish to continue? (If all subjects agree to continue) Thank You. (No subjects withdrew.) During the experiment, (assistant's name) will be sitting at this table. (Point to table.) Your task will be to watch her, to observe her behavior, and to record every occurrance of two specific behaviors, which I will describe shortly.
>
> Whenever you see one of the behaviors occur, you will press the appropriate button on your recording set (point out recording set) and continue to press until the behavior stops. When you press the button, be sure you do it fully. A full press will turn on the light above the button; releasing the button will turn the light off. Anytime you are pressing a button, the light above it should be on. The two behaviors you will be observing will be: (Experimenter will describe the

appropriate two behaviors for that condition) touching the index finger on this side (point to assistant's left side) to the back touchplate on that side, like this (assistant demonstrates) and touching the index finger on this side (point to assistant's right side) to the front touchplate on that side, like this (assistant demonstrates). Only record the behavior if you feel sure it is occurring, that is, press the button when you are sure the finger is actually touching the proper touchplate.

These two buttons are the behavior buttons (point out left two functional buttons) and are labeled with the appropriate behavior. The third button (point to Error Button) marked 'Error' is to be depressed only when you feel that you have made an error in observing or pressing the buttons. If you feel that you have made an error, press the 'Error' button once and release it, like this (demonstrate). When you do this, continue to record what you see. Only press those buttons that correspond exactly to what (assistant's name) is doing at that very moment. Don't record what happened several seconds ago even if you missed it and don't try to second guess her.

This experiment will attempt to determine factors which influence observation. I am recording the actual behaviors that (assistant's name) is doing, through the

thimbles on her fingers. At the same time, I am recording what you are doing, which buttons you are pressing, so I can compare your performance to (assistant's name). I'm not interested in your performance as it relates to you as the average person. Don't worry about how well you are doing, just try to do your best and not make mistakes.

To make sure you've all gotten the idea: First, whenever (assistant's name) touches the front touchplate on this side (point to assistant's right side), you should press the left most button (point to it) and keep it pressed until she stops touching the front touchplate. Now, which button would you press if (assistant's name) touches the back touchplate on this side (point to assistant's left)? Go ahead and press it. (Check to see if all are doing it.) Good.

If you find that the buttons you are pressing don't correspond to what (assistant's name) is doing, just press the error button once and begin recording exactly what she is doing. This error button is very important because we want to know if the average person can recognize their mistakes as they make them. So whenever you make a mistake, press the ERROR button once.

These cards (distribute cards) have the definition of each behavior you are observing written on them. You

will have them in case you forget what to do or get the behaviors mixed up. Does everyone understand what you are supposed to do? Are there any questions? (Answer questions or reexplain instructions.)

First, we will do a 5 minute practice run. Don't start until I say 'Begin'. (Five minute trial run followed by feedback. Correct any wrong recording procedures.)

Now we'll do another five minute practice run to make sure you've got the hang of it. Remember to briefly press the ERROR button when you feel you've made a mistake. Is everyone ready? Don't start until I say 'Begin'. (Second 5-minute trial observation. All subjects should record properly during this trial. If not, correct them as they are recording and continue practice until they record correctly for 5 minutes.) (Three subjects required longer second trial periods. None were longer than 10 minutes.) O.K. Now we're going to begin the data collection phase. Does anyone want to get a drink or take a stretch because we've got a 60-minute session to go? (If yes, give short break.) (No subjects desired a rest period.) Is everyone ready? Start recording when I say begin. (60 minute data collection session.) (After data collection.) O.K., that's the end of the experiment. Now that you've had a chance to experience

the observation procedure, I can give you some more information about the study.

(Talk about the following for about 10 minutes and distribute Human subjects Committee credit forms to subjects.)

1) Uses of systematic observation.

2) Forms which data may take (duration, frequency, etc.).

3) Subjective perception of accuracy vs. actual accuracy vs. operational definition.

4) Problems with methods used to estimate the operational accuracy.

5) Types of data collected in this study, experimental design, hypotheses and preliminary findings.

Data Consolidation and Dependent Variables. During the experiment, the data were recorded on the operation recorder chart paper in such a way that each input from the assistant and subjects corresponded to one track 90cm in length. Thus, for each session, there were 11 tracks of data: two from the assistant (one from each hand) and three from each of the three subjects (one from their recordings of each hand and one from the error recognition pushbutton).

The data in each track were divided into six intervals corresponding to the six 10-minute cycles of the experimental stimuli. Each interval was then divided into 150 units, each corresponding to 4 seconds of time during the experimental session. Two dependent variables were determined using these time unit data: accuracy of observation and agreement between observers.

The accuracy of observation was determined for the data recorded from each subject's observations of the assistant's hands by comparing the subject's track for that hand with the data that was electromechanically recorded from the assistant's hand. Each accuracy was expressed as a percent: the number of units in which the subject's recording of the assistant's behavior was identical to the electromechanical recording of the assistant's hand movements divided by the total number of units (150). Thus, for each cell of the experiment, 36 accuracies were calculated: one for each of the six intervals for each of the two hands that were observed for each of the three subjects.

The agreement between observers was calculated in the same manner as the accuracy of observation. Each subject's observations of one of the assistant's hands was compared, first with one of the other two subject's observations of that hand, then with the remaining subject's observations of that hand. All agreements were expressed as percents: the number of units in which one subject's recording of the assistant's behavior was identical to another subject's recording of the same hand movements divided by the total number of units (150). For each cell of the experiment, 36 agreements were calculated: one for each of the six intervals for each of the two hands that were observed for each of the three possible pairs of subjects. Due to lack of independence between data, the agreement data were combined with accuracy data for analysis (see below).

A third dependent variable, frequency of error recognition, was also determined. Because there was no experimental limit placed on the maximum number of instances a subject might report an error, the frequency of these indications in each 10-minute interval was determined rather than a percentage. Because no distinction was made between the hand for which the error occurred, only 18 error recognition frequencies were determined for each experimental cell: one for each of the six intervals for each of the three subjects.

In order to determine the degree to which interobserver agreement is related to or predicts accuracy of observation, the accuracy and agreement data were combined in the following manner. Each agreement was subtracted from its corresponding accuracy and the absolute value of the difference was determined. These differences were summed across the six intervals for each observer and divided by six (i.e., the mean absolute value difference between agreement and accuracy per interval was determined for each of the two agreements for each observer). These means will be referred to as accuracy-agreement differences. A mean of zero would indicate a one to one correspondence between agreement and accuracy; means that are greater than zero indicate a less than perfect relationship. Since the values the accuracy-agreement differences assumed were limited to between 0 and 100, inclusive, the minimum difference between agreement and accuracy, zero, occured when the accuracy

and agreement were equal; the maximum, 100, occured when either the agreement or the accuracy was at maximum, 100, and the other score was at minimum, 0), these data will be referred to as percentages.

For the purpose of statistical analysis, both the accuracy and the accuracy-agreement difference data were transformed using the arcsin transformation (Winer, 1971).

CHAPTER III

RESULTS

Accuracy Data. A four-way multivariate analysis of variance was performed on the transformed accuracies for the left and right hand data (the accuracy data from each hand constituted a dependent variable). The factors of the analysis were assistant (two levels, separation (two levels), rate (three levels), and interval (six levels). A four-way univariate analysis of variance of the same design as the above four-way multivariate analysis of variance was performed on the accuracy data from each hand (each of the dependent variables used in the multivariate analysis). The multivariate and univariate analysis of variance tables are presented in Tables 1,2, and 3, respectively.

Scheffe's method for comparing means was used to determine significant differences between means in all sources of variance which were significant at $p \leq .10$. Utility indicies (estimates of the proportion of the total variance accounted for by a single source of variance) were calculated for left and right hand data using the method suggested by Gaebelein (in preparation). The utility indicies are presented in Tables 4 and 5.

The multivariate analysis of variance revealed the following significant effects: assistant, rate, interval, assistant by separation by rate interaction, and separation by rate by interval interaction. These significant sources of variance demonstrate the degree to which the accuracy of observation is influenced by factors unrelated to the targets of the observation (i.e., the assistants' finger movements).

The assistant effect is found for the left hand data and marginally for the right hand data (see Tables 2 and 3). For both hands, subjects observing Assistant 1 were more accurate than those observing Assistant 2. The mean accuracies for the subjects observing Assistant 1 were 95.9% (left hand) and 94.6% (right hand) and those for subjects observing Assistant 2 were 92.9% (left hand) and 91.8% (right hand). The mean accuracies for all significant effects may be found in Appendix A. Note that the subjects' accuracies for the left hand are not very different from those for the right hand. (No statistical analyses were performed to determine differences between left and right hand data.)

The rate effect is also found for both the left and right hands. The mean accuracies of the subjects who observed Rate Low-High (in which the assistant's left hand displayed signals at the low rate and her right hand at the high rate) were 94.9% (left hand) and 91.1% (right hand). The mean accuracies of the subjects who observed the Rate Low-Low (both of the assistant's hands displayed signals at the low

rate) were 96.8% (left hand) and 96.3% (right hand). The mean accuracies of the subjects who observed Rate High-High were 91.6% (left hand) and 92.1% (right hand). Note that in the rate conditions in which both of the assistant's hands displayed the same signal rate (High-High and Low-Low) the mean accuracies for each hand do not greatly differ while in the disparate rate condition (Low-High) the mean accuracy for the high signal rate is much less than that for the low signal rate. Scheffe's test shows that for the left hand the mean accuracy of the subjects who observed Rate Low-Low is significantly greater than the mean accuracy of the subjects who observed Rate High-High, $p < .01$. This comparison for the right hand is of the same direction but is marginally significant, $p < .10$. For the left hand, the mean accuracy of the subjects who observed Rate Low-High (left hand displayed the low rate) did not significantly differ from those of the other rate conditions. For the right hand, the mean accuracy of the subjects who observed Rate Low-Low is significantly greater than that of the subjects who observed Rate Low-High (right hand displayed the high signal rate), $p < .05$. Thus, for both hands, the mean accuracies for the subjects who observed the low signal rates (condition Low-Low) were significantly greater than the mean accuracies of the subjects who observed the high signal rates (condition High-High) and the right hand mean accuracies of the subjects

who observed in the High (right hand)-Low (left hand) condition.

The interval effect occurs in both the mean accuracies to the assistants' left and right hands. The mean accuracies to the left hand show a gradual, linear decrease across the six 10-minute intervals from a mean accuracy of 95.7% in the first interval to 93.9% in the last. No comparisons among intervals are significant for the left hand data. The right hand data show a nonsignificant decrease in mean accuracy from the first interval, 94.9%, to the second interval, 93.2%. The decrease in mean accuracy, from the first to the third interval, 92.4%, was the largest found, 2.5%. The fourth and fifth intervals are equal to the third and all three are marginally significantly less than the first, $p < .10$. The sixth interval reveals a nonsignificant increase of accuracy to 93.6%.

The interaction between assistant, rate, and separation is significant for left hand data and marginally significant for right hand data (see Tables 2 and 3). Various aspects of this interaction will be subsequently described. The mean accuracies of the subjects who observed Assistant 2 in the separate condition are similar to the overall results of the rate factor. However, no comparisons among the mean accuracies for the right hand data are significant and, for the left hand data, the mean accuracy of the subjects who observed rate High-High is significantly lower than those of

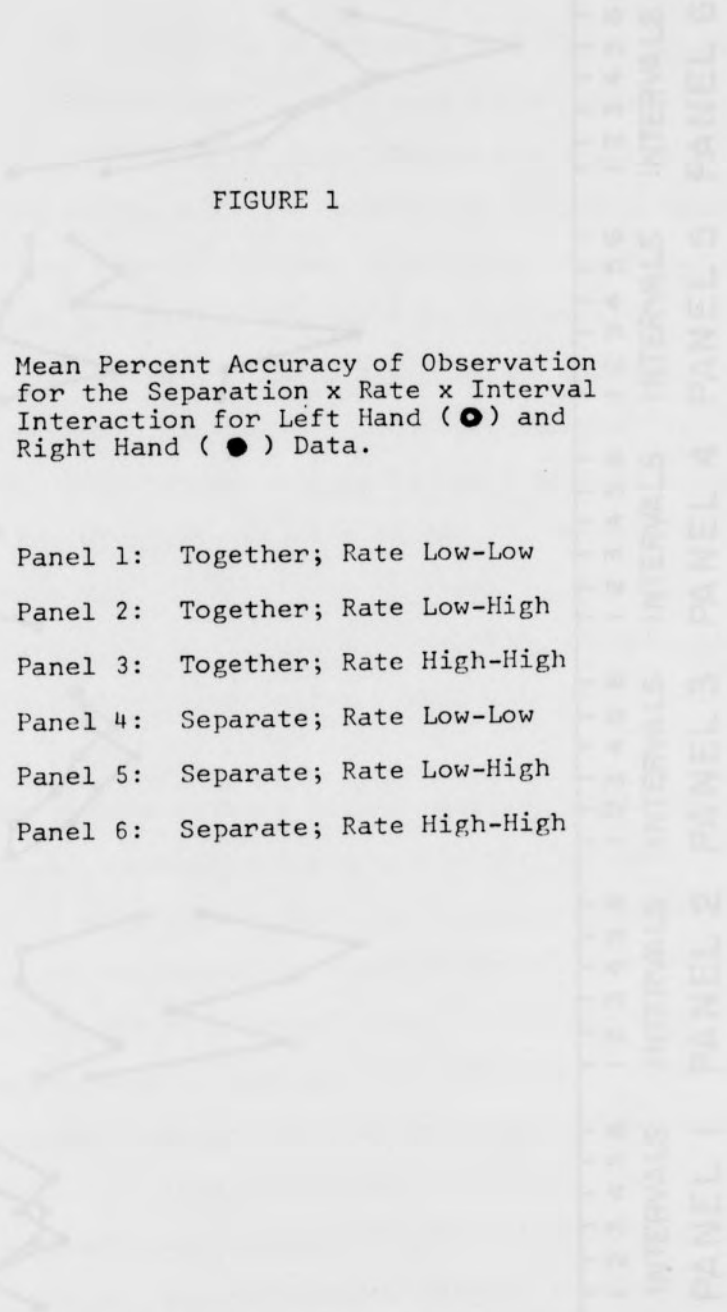
the subjects who observed rate conditions Low-Low and Low-High (left hand displayed the low rate), $p < .01$. The mean accuracies for the left hand data of the subjects who observed Assistant 2 in the separate condition are: 97%, 96.4% and 84.6% for the Low-Low, Low-High and High-High rate conditions, respectively. Note the large differences in accuracy between the Low-Low and Low-High and the High-High conditions.

Large accuracy differences also appear in the right hand data for the subjects who observed Assistant 2 in the together condition. The mean accuracies for these subjects are 98.2%, 86.2% and 90.5% for the Low-Low, Low-High, and High-High rate conditions, respectively. The Scheffe tests show the differences between the Low-Low and the Low-High rate conditions to be significant at $p < .05$ and the difference between the Low-Low and the High-High conditions to be marginally significant at $p < .10$.

The mean accuracies of the subjects who observed Assistant 1 in the together condition show no significant differences between rate conditions. The mean accuracies of the subjects who observed Assistant 1 in the separate condition show marginally significant differences between rate conditions Low-Low and High-High for the data recorded from both hands.

The interaction between separation, rate, and interval is significant for the subjects' observations of both hands. The mean accuracies for this effect are presented in Figure 1.

FIGURE 1



Mean Percent Accuracy of Observation for the Separation x Rate x Interval Interaction for Left Hand (⊙) and Right Hand (●) Data.

Panel 1: Together; Rate Low-Low

Panel 2: Together; Rate Low-High

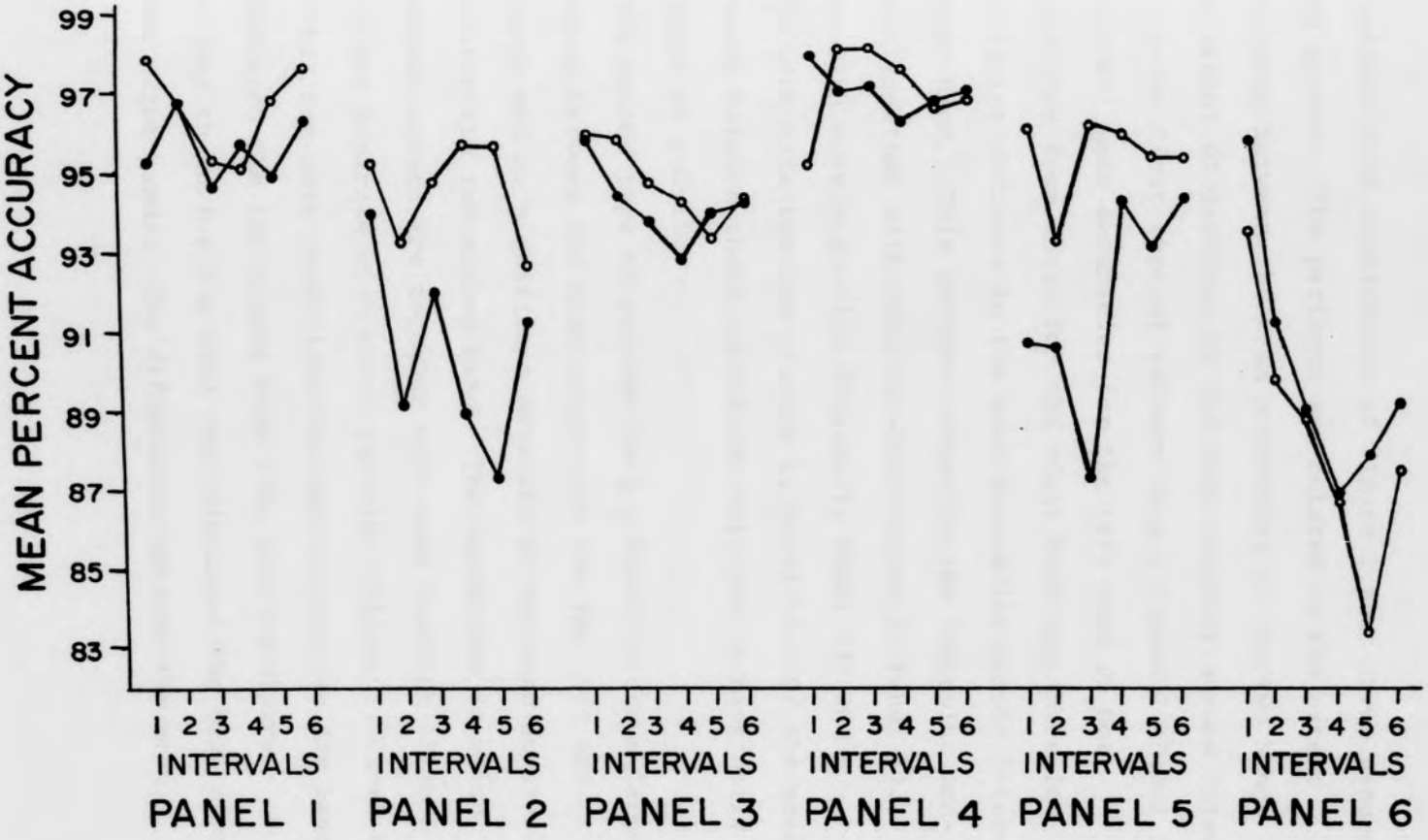
Panel 3: Together; Rate High-High

Panel 4: Separate; Rate Low-Low

Panel 5: Separate; Rate Low-High

Panel 6: Separate; Rate High-High

FIGURE 1


MEAN PERCENT ACCURACY
99
97
95
93
91
89
87
85
83
1 2 3 4 5 6
INTERVALS
PANEL 1
PANEL 2
PANEL 3
PANEL 4
PANEL 5
PANEL 6

It is evident from examination of Figure 1 that three general patterns appear. The patterns are related to the extent of difference between the mean accuracies of the two hands and the extent of decrease in the mean accuracy across intervals. In the first type of pattern (e.g., Figure 1, Panel 1), the subjects' mean accuracies for the left hand do not greatly differ from those for the right hand and there is no significant decrease in the mean accuracies across intervals for either hand. This pattern occurs in the following conditions: together with rate low-low (Figure 1, Panel 1), together with rate high-high (Figure 1, Panel 3), and separate with rate low-Low (Figure 1, Panel 4). Of the means comparisons between these conditions only one is marginally significant at $p < .10$.

The second type of pattern (e.g., Panel 2) shows large differences between the mean accuracies for the left and right hands and no significant decrease in the mean accuracies across intervals for either hand. The conditions in which this pattern occurs are together with rate low-high (Figure 1, Panel 2) and separate with rate low-high (Figure 1, Panel 5). Notice that for both conditions the mean accuracies for hand that displayed the low signal rate (the left hand) are greater than those for the hand that displayed the high signal rate (the right hand). The differences between the mean

accuracies range from 1% in the sixth interval of the separate condition to 8.8% in the third interval of the separate condition. In the other rate conditions (where the displayed rates are equal) large and consistent differences between the mean accuracies for each hand are not found.

For the low-high rate conditions (Figure 1, Panels 2 and 5) there are no significant differences between the mean accuracies of the subjects who observed the assistants' hands separate and those of the subjects who observed the assistants' hands in the together condition. There are six comparisons between mean accuracies that reach or surpass marginal significance when conditions showing Pattern 1 are compared with those showing Pattern 2.

In all of these six comparisons the mean accuracy of the condition showing Pattern 1 is greater than that of the condition showing Pattern 2 regardless of the hand being compared (left or right). However, there are no significant differences between the mean accuracies of the subjects who observed in the together with rate low-high (Figure 1, Panel 2) condition and those of the subjects who observed in the together with rate high-high (Figure 1, Panel 3) condition.

The third pattern appears in the mean accuracies of the subjects who observed the separate with rate high-high condition (Figure 1, Panel 6). There is little difference between

the mean accuracies for the left and right hands but there are significant decreases across intervals for the mean accuracies of both hands. For the left hand, the decrease reaches a maximum of 10.2% at the sixth interval where the mean accuracy is 83.4%. The maximum decrease for the right hand data is 8.8% and occurs in the fourth interval where the mean accuracy is 87.0%. There are also a large number of significant means comparisons between the condition showing Pattern 3 and the other five conditions in the separation x rate x interval interaction. In all of the significant comparisons, the mean accuracy of the separate with rate high-high is the lower of the two.

The utility indicies for the left hand mean accuracy data are presented in Table 4. (Utility indicies are estimates of the proportion of the total variance accounted for by a single source of variance.) The utility indicies for the right hand mean accuracy data are presented in Table 5. Note that of the five significant sources of variance, rate accounts for the greatest proportion of the total variance in both the left (16.6%) and the right (12.2%) hand data. Also note that variance due to the subjects within groups and within subjects across intervals is very high and accounts for 63.6% of the total variance in the left hand data and 74.1% in the right hand data.

Accuracy-Agreement Difference Data. The following data are quite similar and, in some cases, are identical to the accuracy data presented above. They do describe a dependent variable that is separate from accuracy of observation. A three-way multivariate analysis of variance was performed on the transformed accuracy-agreement differences for the left and right hand data where the data for each hand constituted a dependent variable. The factors in the analysis were assistant (two levels), separation (two levels) and rate (three levels). A three-way analysis of variance was performed on the accuracy-agreement difference data from each hand. The multivariate and univariate analysis of variance tables may be found in Tables 6,7, and 8.

Scheffe's method for comparing means was used to determine significant differences between means in all sources of variance which were significant at $p < .10$. Utility indicies were calculated for the univariate analyses of variance and are presented in Tables 9 (left hand data) and 10 (right hand data). The cell means of the accuracy-agreement difference data may be found in Appendix B.

The following effects were found to be significant in the multivariate analysis of variance: assistant, rate, assistant by rate interaction, assistant by separation interaction, and rate by separation interaction, and assistant by rate by separation interaction. These significant sources

of variance show the degree to which the difference between observer accuracy and interobserver agreement is influenced by stimuli that are not related to the targets of observation, the assistants' hand movements.

The assistant effect is found for both the left and right hand data (see Tables 7 and 8). For both hands the accuracy-agreement differences of the subjects who observed Assistant 1 were significantly less than those of the subjects who observed Assistant 2. The mean accuracy-agreement differences for the subjects observing Assistant 1 were 2.4% (left hand) and 3.5% (right hand). The corresponding differences for the subjects observing Assistant 2 were 3.5% and 4.2%.

The rate effect is also found for both the left and right hands. The mean accuracy-agreement differences of the subjects who observed Rate Low-High (assistants' left hand was low rate; right hand, high rate) were 3.0% (left hand) and 4.8% (right hand). The accuracy-agreement differences of the subjects who observed Rate Low-Low were 1.9% (left hand) and 2.5% (right hand); those of the subjects who observed Rate High-High were 4.0% (left hand) and 4.3% (right hand). Note that in the rate conditions in which both of the assistants' hands displayed the same signal rate (High-High and Low-Low conditions), the mean accuracy-agreement differences for each hand do not greatly differ while in

the disparate rate condition (Low-High) the mean accuracy-agreement difference for the high signal rate (right hand) is much greater than that for the low signal rate (left hand). Scheffe's test shows that for the left hand, the mean accuracy-agreement difference for the subjects who observed the low-low rate is marginally significantly less than the mean accuracy-agreement difference of the subjects who observed the low-high rate ($p < .10$) and is also significantly less than that of the subjects who observed the high-high signal rate ($p < .05$). For the right hand data the mean accuracy-agreement differences of the subjects who observed the low-low rate condition were significantly less ($p < .05$) than the mean accuracy-agreement differences of both the subjects who observed Rate Low-High and those who observed Rate High-High. Thus, for both the left and right hand data, the mean accuracy-agreement differences for the subjects who observed in the low-low signal rate condition were significantly less than those of the subjects in the other two rate conditions.

The assistant by rate interaction is significant for both left and right hand data and the assistant by separation effect is significant for left hand data only (see Tables 7 and 8). These effects serve to illustrate the control that aspects of the stimulus display that are a function of the individual being observed exert over the accuracy-agreement difference and will be discussed in more detail in the assistant by separation by rate section below.

In the assistant by rate interaction, for Assistant 1, no means comparisons among rate conditions were significant for either left or right hand data. However, for Assistant 2, the mean accuracy-agreement differences for the subjects observing Rate Low-Low were significantly less than those of the subjects who observed Rates Low-High and High-High for both left and right hand data ($p < .05$).

In the assistant by separation interaction, for Assistant 2, no means comparisons were significant for either left or right hand data. For Assistant 1, the mean accuracy-agreement differences of the subjects who observed the assistant's hands in the together condition were significantly less than those of the subjects who observed in the separate condition ($p < .05$), for the left hand only. No significant differences occurred in means comparisons within the data from the assistant's right hand.

The separation by rate interaction is significant for the data from the left hand ($p < .001$) and is marginally significant for data from the assistants' right hands ($p < .088$). Because of this marginal significance, the results from this interaction must be viewed with caution. For the subjects who observed the assistants' hands in the together position there were no significant differences in mean accuracy-agreement among the rate conditions for the left hand (low-low: 1.9%, low-high: 3.5%, high-high: 2.4%). For the right hand data,

the mean accuracy-agreement differences of the subjects who observed Rates Low-Low (2.6%) and High-High (3.22%) were significantly less than those of the subjects who observed Rate Low-High (5.2%; p .05 and p $<$ .10, respectively). For the subjects who observed the assistants' hands in the separate position the mean accuracy-agreement differences of the left hand data from those subjects who observed in the low-low and low-high conditions were significantly less than those of the subjects who observed in the high-high condition ($p < .05$). The means for the low-low, low-high, and high-high conditions were 1.8%, 2.4%, and 5.6%, respectively. Means comparisons among the right hand data for this condition reveal that the mean accuracy-agreement difference for the subjects who observed the low-low rate condition (2.4%) was marginally significantly less than those of the subjects who observed the low-high (4.4%) and high-high rates (5.2%; $p < .10$). For the data from both hands, the mean accuracy-agreement differences of the subjects observing the high-high rate in the together position were less than those of the subjects observing the high-high rate in the separate position. Thus, for both separation conditions, the mean accuracy-agreement differences of the subjects observing the low-low rate are small for the data from both hands and those of the subjects observing the low-high rate are small for the left hand but are larger for the right hand data. The only

differences between the separation conditions are found in the mean accuracy-agreement differences of the subjects who observed the high-high signal rate. In the together condition, the means are small for the data from both hands while in the separate condition the means are larger for both hands.

The interaction between assistant, rate, and separation is significant for both left and right hand data ($p < .006$). Various aspects of this interaction will be discussed below. The mean accuracy-agreement differences of the left hand data for subjects who observed Assistant 1 in the Together condition were 2.4% (low-low rate), 1.0% (low-high rate), and 1.1% (high-high rate). These means did not significantly differ. The corresponding means of subjects who observed Assistant 2 in the Together condition were 1.5%, 6.1%, and 3.7%. The mean from the low-low rate condition was significantly less than that from the low-high condition ($p < .01$) and marginally significantly less than the mean from the high-high condition ($p < .10$). The mean accuracy-agreement differences of the subjects who observed Assistant 1 in rates Low-High and High were significantly less than the means of the subjects who observed Assistant 2 in these rate conditions ($p < .01$). Alternatively, no significant differences were found between the mean accuracy-agreement differences of the subjects who observed Assistant 1 in the separate position and those who observed Assistant 2 in the

separate position. In these conditions, the left hand data show no significant differences between rates Low-Low and Low-High. For those subjects observing Assistant 2 in the separate position, the mean accuracy-agreement differences at rates Low-Low and Low-High were both significantly less than those of the subjects observing rate High-High ($p < .01$). This comparison is significant only for those subjects observing Assistant 1 at rates Low-Low and High-High ($p < .05$) for both left and right hand data. No rate means comparisons were significant for the right hand data of those subjects observing Assistant 2 in the separate position.

The accuracy-agreement difference data which was discussed above illustrate the extent to which the agreement between observers differs from the observers' accuracy between experimental conditions regardless of the direction of the differences. That is, the calculation of the absolute value of the differences treats occasions in which agreement exceeds accuracy and occasions in which accuracy exceeds agreement in the same manner. This treatment, however, does not permit any understanding of the extent to which the differences occur in both directions. The following section illustrates these differences in a purely descriptive manner. The data are not intended to be representative of any particular condition or of the experiment as a whole; they are the most extreme instances of individual subjects' data.

It was found that 60% of the total number of accuracy-agreement differences involved an accuracy greater than an agreement, 28% involved an accuracy less than an agreement, and 12%, were equal. The accuracy-agreement differences for each interval for each subject in every condition were searched for all differences in excess of 10%. It was found that 7.2% of the total number of accuracy-agreement differences exceeded 10%. Differences in which accuracy exceeded agreement accounted for 5.6% of the total and differences in which agreement exceeded accuracy accounted for 1.6%. The maximum difference in which accuracy exceeded agreement was 26%; the maximum in which agreement exceeded accuracy was 18%.

It is commonly agreed that agreements of 80% or more are acceptable for research. It is thus of interest to determine the extent to which accuracy scores fall above or equal to the 80% criteria while agreements fall below (under estimation of accuracy) and vice versa (overestimation of accuracy). It was found that 3.1% of the total number of differences involve an accuracy of greater than or equal to 80% and an agreement of less than 80%. The opposite situation (agreement $\geq$ 80% and accuracy $<$ 80%) accounts for 2.6% of the total number of differences. It should be noted that 9.3% of all differences where accuracy is less than agreement occur at the 80% criterion while

5% of all differences where accuracy is greater than agreement occur at this level. In contrast, 16.1% of all differences where accuracy is less than agreement occur at the 90% level while 20.6% of all differences where accuracy is greater than agreement occur at this level.

Error Recognition Data. A four way univariate analysis of variance was performed on the frequency of error recognition. The factors of the analysis were assistant (two levels), separation (two levels), rate (three levels), and interval (6 levels). The analysis of variance summary table is presented in Table 11. No sources of variance were significant at $p < .05$. The grand mean was 1.8 errors recognized. The percentage of the total number of errors that were made which were recognized was 9.7%.

By comparing the location of each subject's error indications with that subject's accuracy for the left and right hands, it was found that an error did occur in the subject's recordings of one hand or the other in all but 6.6% of the error indications.

Errors Made by Assistants. The mean number of errors each assistant made for each hand were determined by comparing the operation recorder records of the status of the cue lights with the actual position of the assistants' hands in every 4-second unit of observation. The mean errors made by Assistant 1 on the left and right hands were .8 and 1.2, respectively. Those made by Assistant 2 were 1.67 on each hand.

T-tests were performed on these means between assistants for the data from each hand. Neither T-test was significant at $p < .05$ (left hand: $t(10) = .79$: right: $t(10) = .305$).

TABLE 1

Summary of Multivariate Analysis of Variance of Observer Accuracy (Arcsin Transformation)

Source	Hotelling-Lawley's Trace	Approximate F	df	p
Assistant (A)	.294	3.38	2,23	.050
Separation (S)	.112	1.29	2,23	.294
Rate (R)	1.001	5.51	4,44	.001
A x S	.112	1.29	2,23	.294
A x R	.384	2.04	4,44	.104
S x R	.342	1.88	4,44	.130
A x S x R	.494	2.72	4,44	.041
Interval (I)	.202	2.38	10,236	.011
A x I	.094	1.11	10,236	.355
S x I	.090	1.06	10,236	.391
R x I	.232	1.37	20,236	.138
A x S x I	.112	1.32	10,236	.218
A x R x I	.238	1.40	20,236	.120
S x R x I	.357	2.11	20,236	.005
A x S x R x I	.165	.98	20,236	.508

TABLE 2

Summary of Analysis of Variance of Observer Accuracy for Left Hand Data (Arcsin Transformation)

Source	SS	df	MS	F	p
Assistant (A)	.842	1	.842	7.06	.014
Separation (S)	.254	1	.254	2.13	.158
Rate (R)	1.704	2	.852	7.14	.004
A x S	.233	1	.233	1.96	.175
A x R	.720	2	.360	3.01	.066
S x R	.980	2	.490	4.11	.029
A x S x R	.770	2	.385	3.22	.056
$Error_b$	2.863	24	.119		
Interval (I)	.188	5	.038	2.02	.080
A x I	.128	5	.026	1.37	.240
S x I	.069	5	.014	.74	.596
R x I	.310	10	.031	1.66	.097
A x S x I	.041	5	.008	.44	.821
A x R x I	.416	10	.042	2.23	.020
S x R x I	.378	10	.038	2.02	.036
A x S x R x I	.173	10	.017	.92	.513
$Error_w$	2.239	120	.019		

TABLE 3

Summary of Analysis of Variance of Observer Accuracy for Right Hand Data (Arcsin Transformation)

Source	SS	df	MS	F	p
Assistant (A)	.652	1	.652	3.00	.096
Separation (S)	.026	1	.026	.12	.732
Rate (R)	2.231	2	1.116	5.14	.014
A x S	.011	1	.001	.05	.827
A x R	.839	2	.420	1.93	.165
S x R	.707	2	.354	1.63	.216
A x S x R	1.116	2	.558	2.57	.096
$Error_b$	5.208	24	.217		
Interval (I)	.310	5	.062	3.29	.008
A x I	.066	5	.013	.70	.624
S x I	.097	5	.019	1.03	.402
R x I	.257	10	.026	1.37	.202
A x S x I	.201	5	.041	2.14	.065
A x R x I	.155	10	.015	.82	.607
S x R x I	.462	10	.046	2.45	.011
A x S x R x I	.182	10	.018	.97	.527
$Error_w$	2.257	120	.019		

TABLE 4

Proportion of Variance Accounted for by Sources of Variance in Analysis of Variance for Observer Accuracy (Left Hand Data)

Source	Variance Accounted for by Source
Assistant (A)	.055
Separation (S)	.010
Rate (R)	.112
A x S	.009
A x R	.037
S x R	.057
A x S x R	.040
$Error_b$	.328
Interval (I)	.007
A x I	.003
S x I	-.002
R x I	.009
A x S x I	-.004
A x R x I	.018
S x R x I	.014
A x S x R x I	-.001
$Error_w$	.308
TOTAL	1.000

TABLE 5

Proportion of Variance Accounted for by Sources of Variance in Analysis of Variance for Observer Accuracy (Right Hand Data)

Source	Variance Accounted for by Source
Assistant (A)	.030
Separation (S)	-.013
Rate (R)	.122
A x S	-.014
A x R	.028
S x R	.019
A x S x R	.046
$Error_b$	.465
Interval (I)	.015
A x I	-.002
S x I	.000
R x I	.005
A x S x I	.007
A x R x I	-.002
S x R x I	.019
A x S x R x I	.000
$Error_w$	.276
TOTAL	1.001

TABLE 6

Summary of Multivariate Analysis of Variance of Accuracy-Agreement Difference Data (Arcsin Transformation)

Source	Approximate F (Wilk's Lamba Criterion)	df	p
Assistant (A)	4.16	2,59	.020
Separation (S)	1.68	2,59	.195
Rate (R)	7.21	4,118	.001
A x S	4.00	2,59	.024
A x R	3.40	4,118	.011
S x R	4.75	4,118	.001
A x S x R	7.05	4,118	.001

TABLE 7

Summary of Analysis of Variance of Accuracy-Agreement Differences for Left Hand Data (Arcsin Transformation)

Source	SS	df	MS	F	p
Assistant (A)	.0697	1	.0697	7.47	.008
Separation (S)	.0250	1	.0250	2.68	.106
Rate (R)	.1459	2	.0730	7.82	.001
A x S	.0757	1	.0757	8.11	.006
A x R	.0776	2	.0388	4.15	.020
S x R	.1594	2	.0800	8.54	.001
A x S x R	.1066	2	.0533	5.71	.006
$Error_b$	.5601	60	.0093		

TABLE 8

Summary of Analysis of Variance of Accuracy-Agreement Differences for Right Hand Data (Arcsin Transformation)

Source	SS	df	MS	F	p
Assistant (A)	.0181	1	.0181	1.47	.018
Separation (S)	.0120	1	.0120	.97	.331
Rate (R)	.2139	2	.1070	8.65	.001
A x S	.0005	1	.0005	.04	.840
A x R	.0916	2	.0458	3.70	.030
S x R	.0627	2	.0314	2.53	.088
A x S x R	.2368	2	.1184	9.57	.001
$Error_b$	.7422	60	.0124		

TABLE 9

Proportion of Variance Accounted for by Sources of Variance in Analysis of Variance for Accuracy-Agreement Differences (Left Hand Data)

Source	Variance Accounted for by Source
Assistant (A)	.049
Separation (S)	.013
Rate (R)	.103
A x S	.053
A x R	.048
S x R	.114
A x S x R	.072
$Error_b$	.547
TOTAL	.999

TABLE 10

Proportion of Variance Accounted for by Sources of Variance in Analysis of Variance for Accuracy-Agreement Differences (Right Hand Data)

Source	Variance Accounted for by Source
Assistant (A)	.004
Separation (S)	.000
Rate (R)	.136
A x S	-.008
A x R	.048
S x R	.027
A x S x R	.152
$Error_b$	.640
TOTAL	.999

TABLE 11

Summary of Analysis of Variance of Error Recognition Data

Source	SS	df	MS	F	p
Assistant (A)	17.80	1	17.80	.95	.659
Separation (S)	8.17	1	8.17	.43	.525
Rate (R)	91.18	2	45.59	2.43	.123
A x S	62.30	1	62.30	3.31	.095
A x R	86.79	2	43.39	2.31	.119
S x R	22.75	2	11.38	.60	.561
A x S x R	43.62	2	21.81	1.16	.331
$Error_b$	450.66	24	18.78		
Interval (I)	24.26	5	4.85	2.26	.063
A x I	5.09	5	1.02	.48	.792
S x I	4.94	5	.99	.46	.807
R x I	15.99	10	1.60	.74	.687
A x S x I	13.15	5	2.63	1.22	.303
A x R x I	22.16	10	2.21	1.03	.429
S x R x I	16.97	10	1.69	.79	.639
A x S x R x I	17.43	10	1.74	.81	.620
$Error_w$	257.31	120	2.14		

CHAPTER IV

DISCUSSION

Overview of Results

The results overwhelmingly support the two experimental hypotheses: that interobserver agreement does not necessarily predict the accuracy of observations, and that the accuracy of observation is influenced both by stimuli unrelated to the targets of observation and by those targets, that is, the behaviors the observers have been directed to observe. Only two of the five specific predicted experimental outcomes are supported by the results. These are the existance of large individual differences in the accuracy of recording, and of changes in accuracy across time for most subjects. The following results were not predicted. The existance of a main effect due to the particular assistant being observed was not predicted, but is found in the data. On the other hand, a main effect due to the separation of the assistant's hands was predicted and is not observed. Certain rate effects (e.g., that low rate conditions yield poorer accuracy than high rate conditions) were predicted but other relationships are found. An additional finding was that the subjects did not recognize more than a fraction of their errors. This fraction was uninfluenced by any experimental manipulations.

Relationship Between Interobserver Agreement and Observer Accuracy

The hypothesis that interobserver agreement is not systematically related to observer accuracy is supported by the results of the multivariate and univariate analyses of variance that were performed on the accuracy-agreement difference data. These results show that experimental manipulations can significantly influence the degree to which accuracy of observation can be predicted from interobserver agreement. The assistant being observed, the rate of the behaviors being observed, and interactions of these two variables with themselves and the separation of the behaviors being observed all contribute to the relationship between agreement and accuracy. Individual variables among these significant variables account for up to 15% of the total variance for the right hand data (assistant by separation by rate interaction) and up to 11% for the left hand data (separation by rate interaction). The amount of variance accounted for by the significant sources of variance for the data from each hand were 44% for the left and 36% for the right.

The accuracy-agreement difference variable has the possiblility of yielding three types of data, each type leading to a different course of action. The first type is that in which the accuracy-difference data do not significantly differ from zero. In this case, agreement does not significantly differ from and can be used to predict accuracy in any

of the experimental conditions. The second type is that in which the data significantly differ from zero but this difference remains constant in all experimental conditions. This condition permits prediction of accuracy if the levels of agreement are adjusted by the constant difference. The third type of data, into which class the data from the present experiment fall, yields significant accuracy-agreement differences between experimental conditions and does not permit accuracy to be predicted directly from agreement.

However, it may be possible to empirically determine the correction factor for the constant variables in a particular observation situation (those that have been determined in advance and cannot change during a session; such as separation in the present experiment) and predict accuracies based on agreements obtained from situations with specified parameters. But, even if the parameters of the variables in the observation setting that are constant across time, are known, and can be used to correct the agreement, several problems must first be solved. First, in all analyses, there was a large component of variance accounted for by the variance of subjects within groups and the interaction of the within group variance and intervals of time. This large variance suggests that while group means of accuracy may be accurately predicted from group mean of agreements there is little prediction possible for individual subjects and subjects at a particular moment in time. Second, in

many observation situations, separation and rate variables are not under experimental control: they are situationally dynamic. The experimenter has not set the levels of these variables prior to the observation and has no expectations that they will remain constant. In these situations, there is no way to determine the parameters that are controlling the observer's behavior without relying on the data that is recorded by the observers. Thus, it would often be difficult to determine the information needed to adjust data gathered in a particular observing situation. A third problem that can only be answered through extensive research is the degree to which observers maintain consistant patterns of recording across observation sessions. In other words, to what extent do observers yield equivalent data when moved from experimental to naturalistic situations or even when observing in the same situation on different occasions.

The fourth problem is the significance of the assistant factor and interactions. (This topic is also directly related to the accuracy of observation data and will not be discussed again below.) The experimental assistants were instructed and trained to emit behaviors according to a strict criterion, and the experimenter monitered the assistants' performances during both training and data collection phases in two ways.

The first of these ways, comparing the assistants' actual behaviors to the behaviors programmed for them, (described in Results) revealed no statistical differences between the number of occasions at which the two assistants' respective finger placements did not match those that were programmed. It should be noted that these errors could be perceived only by persons knowing the programmed finger placement. The subjects were not in possession of this information and thus, could not be influenced by the assistants' infrequent errors.

The second way in which the assistants' performances were monitored involved observations that the experimenter made during every session. The experimenter attempted to determine the extent to which differences existed between the assistants' performances with respect to variables.

During training, the assistants were given feedback to reduce these differences that were not mechanically recorded. These variables included the color of clothes worn by the assistant (to eliminate differences in contrast, dark shades of blue, brown, green or grey were used), the actual height of the assistants' fingers (3/4 inch above the touchplate was specified), the angle of the assistants' bodies with respect to the floor and table (perpendicular),and the speed and smoothness with which they changed from one position to another.

Although the experimenter's observations were not as rigorous as those which were electromechanically recorded, once the data collection phase began, he observed no differences between the assistants' performances with regard to the above variables.

The subjects, however, did respond differentially to one or more aspects of the assistants' behavior. It appears that these aspects were not of sufficient magnitude to be noticed by the experimenter even though he received at least 300% more exposure to each assistant than any subject and received exposure to both assistants.

The tentative conclusion drawn from these data and the above considerations is that not only do the situationally dynamic variables (i.e., rate and separation) influence the relationship between agreement and accuracy, but that subtle differences in the topography of the stimuli being presented can also exert this control.

The magnitude and direction of the accuracy-agreement differences are also important parameters. The inability of agreement to predict accuracy would be of little concern if agreements were consistantly lower than accuracy or if the differences were of consistantly low magnitude. However, while most of the differences involve accuracies greater than or equal to agreements, a large proportion (28%) are of the opposite type. This finding suggests that when interobserver agreement is used to estimate accuracy, a sizable proportion of the conclusions made from the agreements will be erroneous.

In terms of magnitude, 7.2% of the differences (both types included) exceed 10%. The maximum difference by which accuracy exceeds agreement is 26%, and by which agreement exceeds accuracy, 18%. This proportion of large differences indicates that in addition to there being a high probability of erroneously assuming accuracy to be at least equal to agreement, there exists a probability that the errors are quite large. This probability, of course, depends upon the experimenter's definition of large (i.e., the point at which discrepancies between accuracy and agreement become unacceptable). The actual cutoff point that is used (arbitrarilly 10% in the present experiment) will depend upon the cost of making an error of that size (falsely assuming accurate data when the accuracy is less than the agreement) and is ultimately a value judgement made by an experimenter. A low cutoff point will decrease the probability of making this type of error but will invalidate more data than would a high cutoff. An experimenter who accepts data containing large differences between accuracy and agreement must demonstrate that the results are of sufficient magnitude or, preferably, are replicable so that the probability that the effect is due only to error can be discounted. On the other hand, the experimenter who accepts only small accuracy-agreement differences must demonstrate that the observation procedure did produce data within that range or else disregard all suspect data.

Variables Influencing Observer Accuracy

The second experimental hypothesis, that the accuracy of observation is influenced by variables in the observation situation that the observers are directed to ignore, is supported by the results of the multivariate and univariate analyses performed on the accuracy data. Each significant effect serves to demonstrate that factors which are not directly targets of the observation do exert influence on the accuracy of the observations. The results emphasize that human observers are controlled as much by extraneous variables as they are by those variables of interest to the experimenter. Certain significant effects deserve individual consideration. The assistant factor was discussed above and need not be covered again.

Effects of Rate. The rate main effect is of great interest for several reasons. The results of the post hoc means comparisons show that accuracy of observation of a behavior (hand movement) at low rate was higher than the accuracy found at high rate. When one behavior occurred at a low rate and the other at a high rate, the accuracy for the low rate behavior exceeded that for the high. These results are not consonant with the previously discussed theory of observing response extinction which would predict lower accuracy for events displayed at low rates (Jerison & Pickett, 1964). However, there are several distinct differences between the

typical vigilance paradigm and the present experiment. The signal rates used in the present research are approximately ten times greater than those used in vigilance research. However, the event rate, which has been shown to be the major determinant of the decreases in accuracy that are typically found in vigilance experiments was slower by about a factor of five. In conjunction with the lower event rate, the signal was not a unitary event but consisted of one signal event (finger position) and three non-signal events. The subjects were required to continuously record the signal's presence or absence rather than record with a single button push and they observed two behaviors (left and right hands) simultaneously. Multiple response requirements were found to eliminate the performance decrement across time (Loeb & Aluisi, 1970), but also reduced overall accuracy. The accuracies found in the present experiment are overall quite high. Thus, the present experiment differs in many ways from the classical vigilance paradigm. These differences permit little generalization of results between the two areas.

The significance of the rate factor (and its interactions) has important implications for much applied research where the ultimate effect of an intervention procedure is to modify the rate of one or more behaviors. Two types of designs illustrate the possible problems: the ABA (return to baseline) and the multiple baseline designs.

In a hypothesized experiment using an ABA design, two behaviors of the same rate are measured by the use of naturalistic observation during a baseline condition. (The measurement of two behaviors in the ABA design is not a requisite of the design, but is necessary in this example because two behaviors were measured in the present experiment.) After a period of time, intervention is begun and the rates of the behaviors are altered. Then the original contingencies are reinstated and observation is continued until the end of that phase. If both behaviors are of low rate during baseline, then increase during intervention and return to the low rate during baseline, one would expect to find differences in the accuracy of the data between baseline and intervention: the baseline data would be more accurate than the intervention data.

In the case of the multiple baseline design, the experimenter attempts to alter the rate of one behavior while holding other behaviors (often recorded simultaneously) constant. The ideal multiple baseline design yields the following results: during the first phase,all behaviors are of the same rate, during each subsequent phase, the intervention alters the frequency of one behavior so that in the second phase one behavior is at a different rate than the other behaviors (which are kept at baseline), and in the third phase, two behaviors are at different rates, etc. The results of the

present experiment show that significant differences in the accuracy of observation occur when behaviors of different rates are observed. Thus, designs that create situations where behaviors which had previously been of equal rate are changed to dissimilar rates may yield data of varying accuracy.

Changes in the accuracy of data in different phases of an experiment can lead to misinterpretations of the data regardless of the design and controls that are used. Assume that an experiment yields baseline data of 10 and 13 (arbitrary units) for Behaviors A and B, and 20 and 14 for these behaviors during an intervention phase. If the experimenter has no collateral measurements from which he can estimate trends in his data, he must rely on the observation data alone for his results. He can assume that his data are accurate within his particular limits. However, if he assumes equal accuracy for the two behaviors, he is ignoring the results of the present experiment that accuracy decreases as the rate of the behavior increases. Thus, the observations of Behavior A during intervention may be less accurate than those made during baseline. But if the data for this behavior are inaccurate, perhaps a portion of the observed increase is due to error. Along the same lines, Behavior B may have actually increased from 13 to 18 but due to the increased probability of inaccuracies associated with higher rates, the intervention observations yielded a score of 14. Thus, it appears that regardless of the observed frequency, the actual frequency can, within limits, be quite different.

There are two basic solutions to this dilemma. The first solution advocates a return to reliability theory. As traditional reliability theory states, the obtained score (an observation in this case) is equal to a true score (the actual event) and some degree of error. This approach would not require the deliniation of actual accuracy and would use statistical procedures, such as the analyses of variance, to determine the relative influence that non-target variables (including the observer) exert on the observed data. Chronbach's generalizability theory utilizes this approach (Chronbach, Gleser, Nanda & Rejaratnam, 1972).

The second solution to the accuracy problem would involve the absolute calibration of observers. That is, prior to data collection, the experimenter would empirically determine the relationship between the accuracy of observation and the important variables that are operative within the observation situation. Experiments akin to the present one can provide these data. Within this approach, after identifying variables that exist in his observation situation, the experimenter would determine the extent to which various levels of each variable influence the accuracy of observation. This procedure is equivalent to the calibration of a sensitive scale. The degree of generalization from the laboratory to the naturalistic situation must also be determined. Once these 'operating characteristics' of the observers have been

deliniated, are within the desired range of accuracy, and are found to generalize to natural situations, the data that are produced by the observations can be assumed to be accurate. That is, data are not collected until their accuracy is known and within acceptable levels.

Neither of these solutions has been used extensively. Each requires a large expenditure of energy and time and returns little information of primary interest to the experimenter. However, the evidence is overwhelming that some method of deliniating the accuracy is necessary. As more researchers and editors of journals become aware of this need, these and other procedures will come into more frequent use.

Effects of Separation. The separation main effect was not found to be significant in the multivariate analysis of variance although this significance was predicted. This factor is, however, significant in combination with the assistant and rate, and with the rate and interval factors. It is apparent from both of these interactions that the separation of the two hands being observed made the observation of both hands more difficult in certain conditions (e.g., Assistant 1, rate High-High) yet the other factors were equally powerful (e.g., accuracy for both hands is low in the Together condition for Assistant 2 at rates Low-High and High-High). In other conditions, the accuracy of the observations of one hand were greater than those of the other hand when the hands

were separated (e.g., Assistant 2, rates Low-High and High-High). The separation between behaviors being observed has a potential for being controlled by the experimenter if the subjects of the observation are restrained to one position but in many naturalistic observations the subjects are free to move about. It may be possible to record these within session changes in separation during the observation session and evaluate the resulting data in terms of the separation change. If changes in the data relate directly to the separation changes, the data may be suspect.

Thus, the separation between the behaviors being observed is found to exert control over the accuracy of observation, especially when it occurs in combination with other variables. These results illustrate the potential complexity inherent in any attempt to determine the influence of the observation situation on accuracy. This complexity is further compounded by the interaction of situational variables, such as the extent of separation aspects of the subjects behavior, such as the assistant separation rate interaction. Although it may be possible to obtain accurate recordings of the important aspects of situational variables, the parameters of the non-target aspects of the subjects' behavior cannot be determined independently of the target behavior.

Temporal Effects Within Sessions. Changes in the accuracy across time were predicted, and the interval effect confirms this prediction. While the large performance decrement that is found in vigilance research was not found, this extreme decrease was not expected due to large differences in the rate and complexity between the present experiment and the vigilance paradigm. Although the decreases across time in the session are of small magnitude, time can be an extremely important factor. The separation by rate by interval interaction (Figure 1) shows this fact.

There are three conditions (Together, Low-Low; Together, High-High; Separate, Low-Low) in which the accuracies for each hand do not differ and there are no significant changes in accuracy across time. In two conditions (Separate, Low-High; Together, Low-High), the accuracies do not change across time but there are large differences between the accuracies for the two hands where those for the left hand (low rate) are always greater than those for the right. The last experimental condition (Separate, High-High) shows a performance decrement for both hands but little difference in accuracies between hands.

The patterns discussed above can be tentatively explained in terms of the frequency and proportions of observing responses that each subject allocated to each target. An observing response is a sequence of behaviors emitted by the observer which permit important environmental stimuli (including the

signal, if it is present) to be received by the observer. Although all subjects were given equal amounts of practice observation using the rates that would be displayed during that session (see Braddeley & Colquhoun, 1969), it may be assumed that in those conditions where one observing response was not sufficient to encompass both target positions (i.e., in separate conditions) the subjects allocated an equal proportion of observing responses to each target. That is, it is assumed that the subjects who observed separated targets either did not discriminate rate differences when they were present or did not adjust their observing responses to match the rates if they did perceive the differing rates. It is assumed however, that even though the observers' perception of the actual signal rate was in error, the observers did emit an observing response rate which was appropriate for their erroneous perception of the signal rate.

The three conditions in which subjects produced the first pattern of accuracy described above had two basic commonalities. The rates for each hand were equal in all three and at most, a moderate overall observing response rate was needed to produce a high level of accuracy. (The overall observing response rate is the number of observing responses that must be emitted per unit time, regardless of the target.)

A low overall observing response rate would suffice for the Together, Low-Low condition where one observing response was sufficient to include both target positions. The overall observing response rate for the Separate, Low-Low condition, where two observing responses are necessary to include both targets, may have been approximately twice that for the Together, Low-Low conditions. This rate is assumed to be of a moderate level and well within the subjects' capabilities. The Together, High-High condition required a higher overall observing response rate but one that was still within the subjects' limits.

The condition that produced the third pattern of accuracy, Separate, High-High, has one major difference from the Together, High-High condition: it requires twice as many observing responses in order to gain the same amount of information about the targets. Therefore, it is assumed that the subjects in the Separate, High-High condition must emit a very high rate of observing responses. Support for these assumptions is provided by the similarities between the trends in these two conditions, although no significant interval effects were found in the Together High-High condition. Both conditions show a decline in accuracy from the first to the fourth or fifth intervals, then an increase to the end of the session. The accuracy in the first intervals do not differ. The only difference is in the magnitude of the decline and recovery with the Separate, High-High group displaying greater decline and recovery. Factors

such as eye fatigue or general decreases in arousal may have been produced by the high overall observing response rate required by the Separate, High-High condition and caused a decrease in the quality or quantity of the observing, decision, and recording responses. Because the subjects' access to information about the passage of time was not controlled, no statement about the final increases may be made.

Both conditions that produced the second pattern of accuracy were low-high rate conditions. As stated above, it is assumed that the subjects who observed these rates could not discriminate the rate differences and that they emitted an overall observing response rate appropriate for the low rate being displayed. At this low observing response rate, many more errors will be made in observing the high rate target behavior than the low rate target behavior. These rate related differences in accuracy are found in these groups: the accuracies to the left hand (low rate) are greater than those to the right hand (high rate). In addition, in several intervals the accuracies to the high rate target behavior are significantly less than those to the same hand in the low-low conditions. However, in these low-high conditions, the explanation remains incomplete. In addition to the decreased accuracy to the high rate behavior, the accuracies to the low rate behavior are significantly less than those found in the low-low conditions. Also, there were no

significant differences between the Together, Low-High and Separate, Low-High conditions. These differences would be predicted in terms of an increased overall observing response rate for the separate condition. These discrepant findings emphasize the hypothetical nature of the explanation. Further research must be done to replicate these data and determine the actual parameters of the observing response.

Individual Subject Differences in Accuracy. The last experimental prediction was of large individual differences in the accuracy of recording. The differences are found in the form of the large components of variance accounted for by subjects within groups and the subjects within groups by interval interaction (see Tables 9 and 10). Thus, any particular individual's observations may be extremely different from those of another individual in the same group. These within group differences may be attributable to the one-shot, short-term nature of the present experiment: no subject received more than 10 minutes of practice and there was little time for improvement in the one hour observation session. The individual differences may diminish with increased training or practice. These data indicate that there is considerable variation of individual observers around the group means. This variation further invalidates the assumption that observers are parallel forms of each other.

Error Recognition by Subjects

The error recognition data show that regardless of the experimental conditions, subjects could identify and report only approximately 10% of the total errors that they made. This measure could, however, be of use in adjusting the data recorded by the observers since approximately 94% of the errors that were reported were actually errors. The experimenter could disregard the data from intervals in which errors were reported and thus increase the accuracy of the data. The procedure would be especially appropriate for codes, such as the Patterson or O'Leary codes, where observations are made in specific intervals. However, one problem inherent in the use of this method is the possibility that there is a direct relationship between the disregarded intervals and some particular category or stimulus array. That is, the errors being recognized and subsequently eliminated, may be nonrepresentative of all errors made in terms of the type of error made or in the type of situation that was being observed when the errors were made. This state of affairs would selectively bias the results of the experiment.

Summary

Interobserver agreement was found to be complexly related to observer accuracy. The extent to which accuracy can be predicted from agreement changes as a function of many variables.

These variables include non-target aspects of observation (e.g., separation and rate) and variables idiosyncratic of the observers themselves (e.g., their responses to subtle differences in signal topography). These findings are in direct opposition to the assumptions currently implicit in the use of interobserver agreement as an index of the 'goodness' of observations (i.e., that observers are parallel forms of each other and high agreement is a sufficient criterion for the use of the data acquired through systematic observation).

In addition, observer accuracy itself was found to be a function of the same types of variables discussed above: the accuracy of observation is influenced by more variables than just those which the investigator instructs the observers to observe. In other words, many variables control the sequence of responses which comprise systematic observation. In addition, the relationships between these variables are extremely complex and can often change within and between single observation sessions or phases of an experiment.

The overall implication of these findings is that most, if not all, systematic observation procedures are of unknown accuracy. The actual levels of accuracy attained by any specific procedure will depend upon a large number of procedural and environmental variables. It is important that this widely used measurement tool be more precisely

evaluated. The most likely result of such evaluation would be the reduction or elimination of numerous aggravating sources of variance in one's data that serve to obscure experimental effects. It is unlikely that such action would eliminate any well replicated findings although the number of borderline or paradoxical results may be reduced.

Although one convenient method for the reduction of error in observation is permitting the observer to denote his errors, this method does not eliminate a significant portion of the errors nor is it sensitive to the variables that influence accuracy. Further research is necessary both to determine the extent to which the above findings are generalizable to other observation procedures and situations and to extend the class of variables known to influence accuracy. Once these variables have been deliniated, Chronbach's method or the method of absolute calibration of the observers may be used to determine the degree of error in any particular data In addition, through the use of these methods, observation procedures may be developed in which the influence of unwanted variables is minimized.

BIBLIOGRAPHY

Adams, J., & Boulter, L. Spatial and temporal uncertainty as determinants of vigilance behavior. Journal of Experimental Psychology, 1964, 67, 127-131.

Arrington, R.E. Some technical aspects of observer reliability as indicated in studies of the "talkies". American Journal of Sociology, 1932, 38, 409-417.

Barker, R.G., & Wright, H.F. One boy's day. New York: Harper & Row, 1951.

Bobbitt, R.A., Gordan, B.N., & Jenson, G.D. The development and application of an observational method: Continuing reliability testing. The Journal of Psychology, 1966, 63, 83-88.

Braddeley, A.D., & Colquhoun, W.P. Signal probability and vigilance: A reappraisal of the 'signal-rate' effect. British Journal of Psychology, 1969, 60, 196-178.

Broadbent, D.E. Some effects of noise on visual performance. Quarterly Journal of Experimental Psychology, 1954, 6, 1-5.

Campbell, D.T., & Fiske, D. Convergent and discriminant validation. Psychological Bulletin, 1959, 56, 81-105.

Cronbach, L.J., Glesser, G.C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley and Sons, 1972.

Eckman, T.A. Reducing the cost of obtaining reliability data in applied settings. In L. King (Chair), Resolving methodological issues in direct behavioral observation in applied settings: Or What's a nice observer like you doing in a place like this. Symposium presented at the 81st American Psychological Association Convention, Montreal, 1973.

Endler, N.S., & Hunt, J.M. Sources of behavioral variance as measured by the S-R Inventory of Anxiousness. Psychological Bulletin, 1966, 65, 336-346.

Gaebelein, J. The Biomedical computer program and the Utility Indicies (Appendix B). Chapter in preparation, 1976.

Gellert, E. Systematic observation: A method in child study. Harvard Educational Review, 1955, 25, 179-195.

Goldfried, M.R., & Sprafkin, J.N. Behavioral personality assessment. Morristown, N.J.: General Learning Press, 1974.

Holland, J.G. Human vigilance. Science, 1958, 128, 61-67.

Jenkins, H. The effect of signal rate on performance in visual monitoring. American Journal of Psychology, 1958, 71, 647-661.

Jerison, H.J., & Pickett, R.M. Vigilance: The importance of the elicited observing rate. Science, 143, 970-971.

Johnson, S.M., & Bolstad, O.D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L.A. Hamerlynck, L.C. Handy, & E.J. Mash (Eds.), Behavior change: Methodology, concepts and practice. Champaign, Ill.: Research Press, 1973.

Jones, R.R., Reid, J.B., & Patterson, G.R. Naturalistic observation in clinical assessment. In P. McReynolds, Advances in psychological assessment (Vol. 3). San Francisco: Jossey-Bass; Pubs., 1975.

Kelley, E.L. Assessment of Human Characteristics. Belmont, California: Brooks/Cole Pub. Co., 1967.

King, G.F., Ehrmann, J.C., Johnson, D.M. Experimental analysis of the reliability of observations of social behavior. Journal of Social Psychology, 1952, 35, 151-160.

Kubany, E.S. & Sloggett, B.B. Coding procedure for teachers. Journal of Applied Behavior Analysis, 1973, 6, 339-344.

Lipinski, D., & Nelson, R.O. Problems in the use of naturalistic observation as a means of behavioral assessment. Behavior Therapy, 1974, 5, 341-351.

Loeb, M., & Alluisi, E.A. Influence of display, task and organismic variables on idicies of monitering behavior. Acta Psychologica, 1970, 34, 343-366.

Loeb, M., & Jeantheau, G. The influence of noxious environmental stimuli on vigilance. Journal of Applied Psychology, 1958, 42, 47-49.

Lovaas, O.I., Freitag, G., Gold, V.J., & Kassorla, I.C. A recording method and observations of behaviors of normal and autistic children in free play settings. Journal of Experimental Child Psychology, 1965, 2, 108-120.

Mackworth, N.H. Experiment 12, (1950). Cited in McGrath, J.J., Harabedian, A., & Buckner, D.M. Review and critique of the literature on vigilance performance. (1959). In Human Factors Research, Inc. (Ed.). Studies of human vigilance-An omnibus of technical reports. Goleta, Calif.: Author, 1968.

Mash, E.J., & McElwee, J.D. Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. Child Development, 1974, 45, 367-377.

Mash, E.J., Terdal, L., & Anderson, K. The response-class matrix: A procedure for recording parent-child interactions. Journal of Consulting and Clinical Psychology, 1973, 40, 163-164.

McGrath, J.J., Harabedian, A., & Buckner, D.N. Review and critique of the literature on vigilance performance (1959). In Human Factors Research, Inc. (Ed.). Studies of human vigilance-An omnibus of technical reports. Goleta, Calif.: Author, 1968.

O'Leary, K.D., & Becker, W.C. Behavior modification of an adjustment class: A token reinforcement program. Exceptional Children, 1967, 33, 637-642.

Olson, E.C. The incidence of nervous habits in children. Journal of Abnormal Psychology, 25, 1930, 75-92.

Olson, W.C., & Cunningham, E. Time sampling techniques. Child Development, 5, 1934, 41-58.

Patterson, G.R., Ray, R.S., Shaw, D.A., & Cobb, J. Manual for coding of family interactions, 1969. Available from ASIS/NAPS c/o Microfiche Publications, 305 E. 46th St., New York, N.Y., 10017. Document #01234.

Preparation of manuscripts. Journal of Applied Behavior Analysis, 1969, 2, 1-2.

Reid, J.B. Reliability assessment of observational data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Reid, J.B., Skindrud, K.D., Taplin, P.S., & Jones, R.R. The role of complexity in the collection and evaluation of observation data. Paper presented at the 81st Americal Psychological Association Convention, Montreal, 1973.

Repp, A., Deitz, D., Boles, S., Deitz, S., & Repp, C. Differences among common methods for calculating interobserver agreement in applied behavioral studies. Journal of Applied Behavior Analysis, in press.

Romanczyk, K.R., Kent, R., Diament, C., & O'Leary, K. Measuring the reliability of observational data: A reactive process. Journal of Applied Behavior Analysis, 1973, 6, 175-184.

Solandt, D.Y., & Partridge, D.M. Research on auditory problems presented by naval operations (1946). Cited in McGrath, J.J., Harabedian, A., & Bucker, D.N. Review and critique of the literature on vigilance performance (1959). In Human Factors Research, Inc. (Ed.). Studies of human vigilance-An omnibus of technical reports. Goleta, Calif: Author, 1968.

Taplin, P.S., & Reid, J.B. Effects of instruction set and experimenter influence on observer reliability. Oregon Research Institute Research Bulletin, 1972, 12 (Whole No. 11).

Taub, H.A., & Osborne, F.H. Effects of signal and stimulus rates on vigilance performance. Journal of Applied Psychology, 1968, 52, 133-138.

Thorne, B.M., Schlottmann, R.S., & Seay, B. Single animal observation versus pair observation in Macaca Irus. Journal of Genetic Psychology, 1969, 115, 17-32.

Winer, B.J. Statistical principles in experimental design. New York: McGraw-Hill Book Co., 1971.

APPENDIX A

Mean Percent Accuracy for all Experimental Conditions for Left Hand Data

Experimental Condition	Interval 1	2	3	4	5	6
Assistant 1						
Together						
L-L*	96.7	95.3	93.8	94.0	96.9	97.5
L-H*	99.6	98.4	98.7	97.3	97.1	97.1
H-H*	97.5	96.9	98.2	98.9	97.3	99.1
Separate						
L-L	93.3	99.1	98.9	97.3	97.3	97.6
L-H	96.0	91.4	94.2	94.6	94.5	96.2
H-H	95.8	91.1	92.2	92.2	89.1	92.2
Assistant 2						
Together						
L-L	98.9	97.8	96.9	96.2	96.7	97.8
L-H	91.1	88.2	90.0	94.0	94.2	86.2
H-H	94.2	94.7	91.3	89.8	89.5	89.5
Separate						
L-L	97.3	97.3	97.6	97.8	95.8	96.0
L-H	96.2	95.3	98.2	97.3	96.4	94.7
H-H	91.3	88.9	85.6	81.1	77.8	82.9

* L=Low Rate Condition
H=High Rate Condition

APPENDIX A

Mean Percent Accuracy for all Experimental Conditions for Left Hand Data

Experimental Condition	Interval 1	2	3	4	5	6
Assistant 1						
Together						
L-L*	96.7	95.3	93.8	94.0	96.9	97.5
L-H*	99.6	98.4	98.7	97.3	97.1	97.1
H-H*	97.5	96.9	98.2	98.9	97.3	99.1
Separate						
L-L	93.3	99.1	98.9	97.3	97.3	97.6
L-H	96.0	91.4	94.2	94.6	94.5	96.2
H-H	95.8	91.1	92.2	92.2	89.1	92.2
Assistant 2						
Together						
L-L	98.9	97.8	96.9	96.2	96.7	97.8
L-H	91.1	88.2	90.0	94.0	94.2	86.2
H-H	94.2	94.7	91.3	89.8	89.5	89.5
Separate						
L-L	97.3	97.3	97.6	97.8	95.8	96.0
L-H	96.2	95.3	98.2	97.3	96.4	94.7
H-H	91.3	88.9	85.6	81.1	77.8	82.9

* L=Low Rate Condition
H=High Rate Condition

APPENDIX A
(Cont.)

Mean Percent Accuracy for all Experimental Conditions for Right Hand Data

Experimental Condition	Interval 1	2	3	4	5	6
Assistant 1						
Together						
L-L*	92.2	93.6	90.9	94.2	92.2	94.6
L-H*	96.7	94.7	94.7	93.5	91.5	95.3
H-H	96.9	98.6	98.0	98.0	97.3	98.2
Separate						
L-L	98.5	98.0	98.4	96.0	99.1	96.9
L-H	93.6	92.7	92.5	96.7	93.6	94.9
H-H	95.1	91.1	91.1	85.8	89.8	90.0
Assistant 2						
Together						
L-L	98.4	99.6	98.4	97.4	97.8	98.0
L-H	91.3	83.8	89.3	84.4	83.3	85.3
H-H	94.7	90.2	89.6	87.5	90.7	90.2
Separate						
L-L	97.6	96.2	96.0	96.7	94.4	97.1
L-H	88.0	88.7	82.4	92.0	92.9	94.0
H-H	96.4	91.6	87.3	88.2	86.2	88.6

* L=Low Rate Condition
H=High Rate Condition

APPENDIX B

Mean Percent Accuracy-Agreement Difference Data

Experimental Condition	Left Hand	Right Hand
Assistant 1		
Together		
L-L*	2.37	4.26
L-H*	.95	4.37
H-H*	1.06	1.29
Separate		
L-L	1.94	1.81
L-H	2.93	3.79
H-H	5.06	5.66
Assistant 2		
Together		
L-L	1.48	1.00
L-H	6.10	6.08
H-H	3.74	5.34
Separate		
L-L	1.71	3.03
L-H	1.96	5.00
H-H	6.18	4.78

* L=Low Rate Condition
H=High Rate Condition